RANDOM DOT PRODUCT GRAPHS

A MODEL FOR SOCIAL NETWORKS

by

Christine Leigh Myers Nickel

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

December, 2006

Abstract

We develop a new set of random graph models. The motivation for these models comes from social networks and is based on the idea of common interest. We represent a social network as a graph, in which vertices correspond to individuals. In the general model, an interest vector x_v, drawn from a specific distribution, is associated with corresponding vertex v. The edge between vertices u and v exists with some probability $P(x_i, x_j) = f(x_u \cdot x_v)$; that is, it is equal to a function of the dot product of the vectors. The probability of a graph H is given by $P_{\mathrm{X}}[H] = \left[\prod_{\substack{uv \in E(H)\\ u<v}} f(x_u \cdot x_v)\right] \left[\prod_{\substack{uv \notin E(H)\\ u<v}} (1 - f(x_u \cdot x_v))\right]$ and is dependent upon the distribution from which the vectors are drawn.

We examine three versions of the Random Dot Product Graph on n vertices. In the dense model, the vectors are drawn from $\mathcal{U}^a[0,1]$, the a^{th} power $(a > 1)$ of the uniform distribution on $[0,1]$, and f is the identity function. In this case, with probability approaching one as n approaches infinity, all fixed graphs appear as subgraphs. In the sparse model, the vectors are again drawn from $\mathcal{U}^a[0,1]$, however the probability function is $f(s) = \frac{s}{n^b}$ $(b \in (0, \infty))$. With this change, subgraphs appear at certain thresholds and we examine the sequence of their appearance. In both cases, we show that the models obey a power law degree distribution, exhibit clustering, and have a low diameter; these are all characteristics found in social networks.

The third version is a discrete model. Here the vectors are drawn from $\{0,1\}^t$ ($t \in \mathbb{Z}_+$)) and $f(s) = \frac{s}{t}$. Each coordinate of x_v is independently assigned the value 1 with probability p and 0 otherwise ($p \in [0,1]$). We define the probability order polynomial, or POP, of a graph H as a function that is asymptotic to $P_{\geq}[H]$, the probability of H appearing as a (not necessarily induced) subgraph, and present geometric techniques for studying POP asymptotics. We give a general method for calculating the POP of H. We present formulas for the POPs and first moment results for trees, cycles, and complete graphs. We also prove a threshold result for K_3 and describe a general method for proving threshold results when all the required POPs are known.

Advisor: Edward R. Scheinerman

Second Reader: Carey E. Priebe

Acknowledgements

This research was supported in part by the Australian Defence Science Technology Organisation; Miro Kraetzel, project officer.

I cannot possibly express in words the gratitude that I owe my advisor, Edward Scheinerman. He is a brilliant mathematician and most excellent mentor. In addition to his insights and guidance, I greatly appreciate his understanding and patience. I am most thankful that he never gave up on me.

I would like to thank Professors Priebe, Wierman, and Fishkind for serving on my dissertation committee. Not only did they offer suggestions on improving this thesis, but they asked insightful questions and showed a true interest in this work.

In addition to my committee, Miro Kraetzel and Professor James Fill helped to improve this work with their ideas and suggestions.

I am thankful to all of the faculty who taught or otherwise supported me during my graduate schooling, both here at Johns Hopkins and also at the University of Alabama in Huntsville. I would like to specifically thank Peter Slater for being my earliest advisor and Grant Zhang for introducing me to Graph Theory.

I would like to thank the current and former staff of our department, Amy Berdann, Kristin Bechtel, Sandy Kirt, and Kay Lutz, for all their support.

I found many friends amongst my fellow graduate students and I am grateful for them all. I would like to thank Leslie Cope for that very fist sentence. I am especially thankful for Beryl Castello. She has been the best baby sitter, cheerleader, and friend.

I owe a deep gratitude to my family. My parents, Norm and Maryellen, and siblings, Jay and Jenny, have always believed in and supported me; not just in this endeavor, but in all that I do.

Finally, I am eternally thankful for my husband Lee and daughter Alexandra. They have made many sacrifices so that I could pursue my dream, always doing so without complaint. Lee has been the best husband possible. He has loved and supported me in every way. He has always believed in me and I am so thankful that he never let me quit. Lee's love and confidence in me and Alexandra's smile made this dissertation a reality.

This dissertation is dedicated to all of my children: those who have gone before me, those who are with me now, and those who I have yet to meet.

Contents

Abstract **ii**

Acknowledgements **iv**

List of Figures **ix**

1 Introduction **1**
1.1 The Motivating Problem 1
1.2 A Brief History of Modeling the Internet 2
1.2.1 Early Random Graph Models 2
1.2.2 The Internet and Power Laws 5
1.2.3 Power Law Generators 7
1.2.4 The Internet as a Small World 10
1.3 An Overview of Things to Come 11

2 The Dense Model **15**
2.1 The Random Dot Product Graph – A New Model 15
2.1.1 The General Model 15
2.2 The Dense Random Dot Product Graph 17
2.2.1 Some Results in the Simplest Case 17
2.3 Main Results 21
2.3.1 Degree Power Law 21
2.3.2 Clustering 38
2.3.3 Diameter 39
2.4 Higher Dimensions 47
2.4.1 Basic Results in Higher Dimensions 47
2.4.2 A Bend in the Power Law 51

3 The Sparse Model **56**
3.1 The Sparse Random Dot Product Graph 57

3.1.1 Results in the General Case: $b \in (0, \infty)$ 57
3.1.2 Results when $b \in (0, 1)$ 69
3.1.3 The Land of Trees: $b \in (1, 2)$ 78
3.2 Main Results 87
3.2.1 The Threshold Result 87
3.2.2 Degree Distribution 92
3.2.3 Diameter 110
3.3 The Complete Picture 119

4 Posynomial Asymptotics: A Prelude to the The Discrete Model 122
4.1 The β, α Parameter Space 122
4.2 The Equivalence Classes 127
4.3 Dividing Lines 138

5 The Discrete Model 153
5.1 The 0, 1 Discrete Random Dot Product Graph 154
5.1.1 Basic Results 155
5.1.2 $\mathbf{K_3}$: A Detailed Example 163
5.2 The Probability Order Polynomial 168
5.2.1 Defining the POP 168
5.2.2 POPs of Trees, Cycles, and Complete Graphs 177
5.3 Towards a Threshold Result 196
5.3.1 The First Moment 196
5.3.2 Threshold Theory 214

6 Overview and Directions for Future Work 231
6.1 Observations and New Directions 232
6.1.1 The Dense Random Dot Product Graph 232
6.1.2 The Sparse Random Dot Product Graph 234
6.1.3 The Discrete Random Dot Product Graph 237
6.2 Other Open Problems and Thoughts for Future Work 240
6.2.1 General Graph Theory Questions 240
6.2.2 Variations on the Model 241
6.2.3 In Conclusion 242

Bibliography 244

Vita 248

List of Figures

2.1 The domain $\mathcal{U}(b)$ 53

4.1 The line $l_i : h - x_i\alpha - \beta\, i = 0$. 124
4.2 **(a)** The lines associated with $f_{\alpha,\beta,h}(n)$. **(b)** $\mathcal{C}(f)$ and $\zeta(f)$. 126
4.3 If $(\beta_0, \alpha_0) \in \mathcal{C}(f)$ then all points $(\beta, \alpha) \geq (\beta_0, \alpha_0)$ are in $\mathcal{C}(f)$. 128
4.4 The lines l_i, $l_{i,j}$ and l_j. 137
4.5 Case 1: The region A. 142
4.6 Case 1 with **(a)** $x_i > y_j$ and the slope of l_3 positive. **(b)** $x_i < y_j$ and the slope of l_3 negative. 144
4.7 Case 2: The region A'. 145
4.8 Case 3: The region A''. 146
4.9 Case 3 with $\beta_{int} < 0$ and **(a)** $x_i > y_j$. **(b)** $x_i = y_j$. **(c)** $x_i < y_j$ and $i < j$. **(d)** $x_i < y_j$ and $i = j$. **(e)** $x_i < y_j$ and $i > j$. 147
4.10 Case 4: The region A''. 149

5.1 **(a)** The lines that define $\mathcal{C}(f'^{T})$. **(b)** l_i intersects $\mathcal{C}(f^T)$ at exactly one point. 199
5.2 **(a)** The lines that define $\mathcal{C}(f'^{C_h})$. **(b)** l_i is 'below' $\mathcal{C}(f^{C_h})$. 203
5.3 The lines that define $\mathcal{C}(f'^{K_h})$. 207
5.4 The line l_k is 'below' $\mathcal{C}(f^{C_h})$. 210
5.5 The graph H_{kite}. 211
5.6 **(a)** All of the lines associated with terms of $f^{H_{kite}}$. **(b)** $\mathcal{C}(f^{H_{kite}})$ and $\zeta(f^{H_{kite}})$. 213
5.7 The dotted lines are those associated with terms of $g^{*}_{\alpha,\beta,4}(n)$ and are all 'below' $\mathcal{C}(f'^{K_3})$. 219
5.8 The dotted lines are those associated with terms of $g^{**}_{\alpha,\beta,5}(n)$ and are all 'below' $\mathcal{C}(f^{K_3})$. 220
5.9 The graph H_{2-kite}. 227
5.10 The line $l(H_{2-kite})_0$ is not always 'below' $\mathcal{C}(f^{H_{kite}})$. 228

5.11 **(a)** The point $(7/10, 10/29) \in \mathcal{C}(f^{H_{kite}})$. **(b)** The point $(7/10, 10/29) \in \zeta(f^{K_4})$. 229

Chapter 1

Introduction

In this thesis we develop a new set of random graph models. The motivation for these models comes from social networks and is based on the idea of common interest.

1.1 The Motivating Problem

People with shared interest are more likely to communicate than those without shared interests. Perhaps an active graph theorist is more likely to interact with another discrete mathematician than with a statistician. Likewise, a drastic change in an individual's social relationships may reflect a corresponding change in that individual's lifestyle or activities. If our graph theorist suddenly begins to socialize with literary critics, we might expect that this was due to a change in the mathematician's interests.

We represent a social network as a graph, in which vertices correspond to individuals. An interest vector x_i is associated with each individual and corresponding vertex i. The edge between vertices i and j (indicating acquaintance between the corresponding individuals) exists with some probability $P(x_i, x_j)$. The internet may be modeled in the same way: Individual web spages are represented by vertices with (directed) edges corresponding to hyperlinks.

We seek to develop tractable and realistic models for such graphs. A good probabilistic model would lead to simulated networks whose graph theoretic characteristics match those of naturally occurring networks. Several distinct graph theoretic features are repeatedly observed in social networks; for example the degree distribution typically follows a power law and the network appears to be highly clustered.

1.2 A Brief History of Modeling the Internet

The world wide web is a particularly attractive subject for network research. Although it is a large network, techniques have been developed to collect information about its structure. Additionally, the web is dynamic, changing continuously. Finally, unlike other social networks, data collection may be done anonymously.

1.2.1 Early Random Graph Models

Originally attempts were made to model the physical structure of the internet. The internet was viewed as an undirected graph in which nodes represented routers (or domains) and edges represented the physical links connecting the routers (domains).

An early random graph model for the internet was proposed by Waxman in 1988 [24]. We distribute N vertices uniformly at random in the x-y coordinate plane. For each pair of vertices, u and v, the Euclidean distance, $d(u,v)$, is calculated and a parameter L is defined to be the maximum value of these distances. Two additional parameters, $\alpha, \beta \in [0,1]$, are chosen by the user. The probability that the edge between u and v is in the graph is $P[u \sim v] = \beta\, e^{\frac{-d(u,v)}{L\alpha}}$. Once all possible vertex pairs have been considered the graph is checked for connectedness. If the graph is found to be disconnected, then process restarts and repeats itself until a connected graph is generated.

In 1993, Doar and Leslie [9] proposed a modification to the Waxman graph that would allow the average degree of a vertex to remain fixed as the number nodes increased. Again, N nodes are distributed uniformly at random in the x-y plane and the Euclidean distances $d(u,v)$ for each pair of vertices u and v are calculated. However, in addition to the parameters L, α, and β of the Waxman graphs, the desired average degree ξ, and a constant k are introduced. The constant k varies with α, β, and ξ and is determined numerically. The probability that an edge exists between vertices u and v is given by $P[u \sim v] = \beta\, \frac{k\xi}{N}\, e^{\frac{-d(u,v)}{L\alpha}}$.

Calvert et al. [8] sought a model that more closely reflects the hierarchical structure of the internet. Their model consists of a two phase process. First, N vertices are uniformly distributed in the unit square and the Euclidean distances between each pair of vertices u and v, $d(u, v)$ is calculated. A radius R and probability α are chosen by the user so that whenever $d(u, v) \leq R$, the vertices u and v are adjacent with constant probability α. If the distance $d(u, v)$ is above the threshold R, the probability of the edge between u and v decreases linearly as $d(u, v)$ increases. The probability that an edge exists between u and v is given by

$$P[u \sim v] = \begin{cases} \alpha & \text{if } d(u,v) \leq R, \\ \alpha \frac{\sqrt{2}-d(u,v)}{\sqrt{2}-R} & \text{if } d(u,v) > R. \end{cases}$$

If the resulting graph is not connected, the process starts over. Once a connected graph G has been generated the second phase begins.

In the second phase, the graph G determines the upper-level structure for the final graph. Each of the N nodes in the graph G is replaced by another graph on N vertices that is generated in the same manner as G. Edges incident to a node u in G are connected to vertices in the new graph G_u that replaced u in G as follows: order the vertices of G_u in increasing order of degree ignoring all vertices of degree 1; connect the edges in the upper-level graph that are adjacent to u in G to the vertices in G_u, one edge per vertex, sequentially with respect to the above ordering. The resulting graph will have N^2 vertices. The process can be easily generalized to create a graph with multiple levels.

Another hierarchical generation method is the Transit-Stub method introduced by Calvert, Doar and Zegura [7]. The basic idea is that a connected random graph is generated using any of the above methods. Then each vertex in the graph represents a transit domain and is replaced by another randomly generated, connected graph G_{td}. Then, for each vertex in a transit domain graph G_{td}, a collection of random graphs is generated and attached to its transit domain vertex. These graphs represent the stub domains. Finally, extra edges are added between stub domains or stub domains and transit domains. This generates a random graph with the desired hierarchy.

Other models have been proposed [25], however they mainly consist of adaptations to these basic ideas. Additionally, canonical networks such as the rings, paths, the rectangular grid and the Erdős-Rényi random graphs have been studied, but deemed unrealistic for this application [21, 25].

1.2.2 The Internet and Power Laws

In August of 1999, Faloutsos et al. released a technical report that changed the way in which modelers viewed the internet [10]. They looked at three domain level "snapshots" of the physical (interdomain) topology of the internet taken during the period from November 1997 to December 1998. During this time the internet experienced 45% growth. They concluded that internet exhibited at least 3 different power law relationships.

The first of these power laws relates the out-degree d_v of a vertex v to the rank r_v of the vertex, where a vertex corresponds to an internet domain. The rank of a vertex v is the index of v when the vertices are listed in decreasing order of out-degree.

They showed that when all vertices of out-degree zero are ignored, $d_v \propto r_v^R$ where R, the rank exponent, is the slope of the log-log plot of (r_v, d_v). The next power law is the degree distribution power law. This law illustrates the relationship between the frequency f_d of the out degree d to d. Again, ignoring vertices of out-degree zero, it was found that $f_d \propto d^{\phi}$ where the out-degree exponent ϕ is the slope of the log-log plot of (f_d, d) and is negative. Faloutsos et al. felt that this power law was the most important since the data supported it most closely. A third power law relates the positive eigenvalues of the adjacency matrix of the internet graphs to their index when listed in decreasing order, i.e., $\lambda_i \propto i^{\varepsilon}$.

Power laws are also exhibited elsewhere in the World Wide Web structure. If the web is modelled as a directed graph in which vertices represent documents and edges represent hyperlinks from one document to another, then Albert et al. [3] found that the frequency f_d of an out-degree d was again inversely proportional to a power of the degree d and that this relationship is a power law. Independently, Kleinberg et al. obtained the same results [15]. Huberman and Adamic also found that the number of web pages at each of the web sites exhibits a power law behavior similar to the Faloutsos rank versus out-degree relationship [12]. These results shifted interest in the modeling community away from hierarchy and towards degree distribution.

1.2.3 Power Law Generators

Although Faloutsos et al. [10] discovered power laws inherent in the internet, they did not attempt to explain why the behavior occurred. Two basic causes for the degree distribution power law were suggested by Barabási and Albert [4]. The first is that the web exhibits incremental growth, that is the size of the network is gradually increased over time by the continual addition of new vertices. The second idea is that of preferential connectivity, or the increased likelihood that a new vertex will be adjacent to an existing vertex of higher degree.

Medina, Matta, and Byers [16] proposed a parameterized topology generator that allows the user to study the causes suggested by Barabási and Albert. Their model BRITE (for Boston University Representative Internet Topology Generator) divides the x-y plane into equally sized "high-level" squares. Each high-level square is subdivided into equally sized "low-level" squares. The number of vertices n to be placed in each high-level square is determined by a specific distribution. The n vertices are then distributed uniformly in the high-level square ensuring that at most one vertex is placed in each low-level square. BRITE has three parameters that control how vertices are connected. The parameter m determines the number of neighbors to which a new vertex will be connected when it first joins the graph. The incremental growth parameter can either be INACTIVE: that is, all of the vertices will be placed in the plane before any of the edges are added and then at each step a vertex is randomly selected and attached to m other vertices in the plane; or ACTIVE: here a small randomly connected backbone of m_0 vertices is initially generated and then all other vertices are added one at a time and connected to vertices that are already in the graph. Finally, the preferential connectivity parameter controls the probability that two vertices will be adjacent. If the preferential connectivity is set to NONE: then the Waxman probability function is used; if set to ONLY: then a vertex v, when first considered, will connect to a possible neighbor u with probability $P[v \sim u] = \frac{d_u}{\sum_{x \in C} d_x}$ where d_u is the degree of vertex u, and C is the collection of possible neighbors; if set to BOTH: then the vertices are adjacent based on a combination of Waxman's probability $w_{v \sim u}$ and the above preference for already highly connected nodes, so that $P[u \sim v] = \frac{w_u \, d_u}{\sum_{x \in C} w_x \, d_x}$. They found that the preferential connectivity and incremental growth were both needed to ensure that the graphs generated obey the degree distribution power law.

Another model for producing graphs that obey the degree distribution power law was proposed by Aiello, Chung, and Lu [2]. Their model has two parameters, α and β, that control the size and growth rate of the graph. They generate a random graph with a degree distribution that obeys the following: For a given degree d, the number of vertices f_d of degree d is given by $f_d = \lfloor \frac{e^\alpha}{d^\beta} \rfloor$. They assume no isolated vertices. To generate their graphs for each vertex v, they create a set L_v containing $d(v)$ copies of v. Next, they find a random matching on the vertices in $L = \bigcup_v L_v$. Finally, they collapse each L_v into a single vertex so that for any two vertices u and v, the number of edges between u and v is equal to the number of edges in the matching between vertices in L_u and vertices in L_v.

A third generator was developed by Jin, Chen, and Jamin [13]. The generator uses the desired number of vertices N and the percentage k of the vertices that are of degree one to calculate the degree and rank distributions based on the Faloutsos power laws. Next a spanning tree is generated among all the vertices of degree at least two, by beginning with the empty graph G and uniformly at random selecting a node not yet in G to be added. The new vertex v will be adjacent to an existing vertex u in G, that is not already adjacent to $d(u)$ vertices, with probability proportional to the to the current degree of u in G. Next the nodes of degree one are added to G in the same manner. Finally, any remaining vertices with degrees that have not been satisfied are connected, beginning with the vertex of highest degree, with the same proportional probabilities. A feasibility check is used to ensure that the graph is connected.

Additionally, in a vein similar to ours, others have attempted to model social networks such as the internet by assigning vertices to points in space and determining their adjacency based on these positions. Hoff, Raftery, and Handcock [11] propose such a latent position model. In their model, they allow differing levels of adjacency between vertices and represent the social network as a matrix Y with entries $y_{u,v}$ indicated the level of affinity between actors u and v. Thus, one can think of Y as the adjacency matrix representation of a graph with weighted edges. For a fixed Y, the probability of a realization of Y is given by $P(Y|Z, X, \theta) = \prod_{u \neq v} P(y_{u,v}|z_u, z_v, x_{u,v}, \theta)$

and is dependent upon Z, the positions associated with the vertices, X, additional covariate information, and θ, a model parameter. Hoff et al. discuss two specific latent position models. In both models $\theta = [\alpha, \beta]$. In the first model, the probability of the adjacency $y_{u,v}$ depends on the Euclidean distance between z_u and z_v (the positions associated with u and v), specifically, the probability of $y_{u,v}$ depends on $\alpha + \beta \cdot x_{u,v} - |z_u - z_v|$. In their second model, the probability of the adjacency $y_{u,v}$ depends upon $z_u \cdot z_v / |z_v|$, the signed magnitude of the projection of z_u in the direction of z_v, specifically, $\alpha + \beta \cdot x_{u,v} + \frac{z_u \cdot z_v}{|z_v|}$.

Other models have been suggested [5, 6, 18], however they are similar to those listed above. There has been some work done comparing the different random graph based models for the internet [6, 21, 25]. Several metrics that have been suggested include diameter, number of biconnected components, eccentricity distribution, k-neighborhood size distribution, and clustering coefficients.

1.2.4 The Internet as a Small World

The notion of small world phenomenon, or the idea that any two people can be related through a relatively short chain of acquaintances, has been present in social theory for decades [17]. However the idea of small world graph was introduced by Watts and Strogatz [23]. Some basic definitions are needed to describe a small world graph [22]. Consider a graph $G = (V, E)$. For each vertex $v \in V$ let $\rho(v) = \frac{\sum_{u \in V} d(u,v)}{|V|-1}$ where $d(u, v)$ is the number of edges in a shortest path between u and v. Then $\rho(v)$

is the average shortest path length from v to any other vertex. Let the characteristic path length, L_G be the median of the $\rho(v)$'s. Next we define the clustering coefficient $\gamma(v)$ for each $v \in V$ as

$$\gamma(v) = \frac{|E(N(v))|}{\begin{pmatrix} |N(v)| \\ 2 \end{pmatrix}}$$

where $N(v)$ is the open neighborhood of v. Then the clustering coefficient γ_G for the graph G is the mean of $\{\gamma(v) : |N(v)| \geq 2\}$. Finally, let $R(n, k)$ be the random graph on n vertices with average degree k constructed by beginning with the empty graph on n vertices and uniformly at random choosing two vertices to connect with an edge until the graph contains $\frac{nk}{2}$ edges. Then a small world graph is a graph G on n vertices with average degree k for which $L_G \approx L_{R(n,k)} \approx \frac{\log(n)}{\log(k)}$, but $\gamma_G \gg \gamma_{R(n,k)} \approx \frac{k}{n}$, that is a graph for which the characteristic path length is small and similar to that of a random graph, but whose vertices are highly clustered.

It has been noted that the world wide web exhibits small word behavior [1, 6].

1.3 An Overview of Things to Come

This thesis introduces and examines the Random Dot Product Graph suite of models. Three sets of models for social networks, all based on the idea of common or shared interest are discussed. The first model, discussed in Chapter 2, is the Dense Random Dot Product Graph. In this model, with probability approaching one as the number of vertices approaches infinity, all small subgraphs appear. The second model, discussed in Chapter 3, is the Sparse Random Dot Product Graph and small subgraphs appear at certain thresholds. Finally, the third model is a discrete version and is discussed in Chapter 5. As in the sparse model, in the Discrete Random Dot Product Graph small subgraphs appear at various thresholds.

More specifically, Chapter 2 introduces the general Random Dot Product Graph. We begin by defining and discussing the basic model. We then focus on the Dense Random Dot Product Graph, starting by presenting simple results. The bulk of the chapter is spent proving the three main results in the dense case. First in Section 2.3.1, we show that the model obeys the degree distribution power law observed in World Wide Web [10]. Secondly in Section 2.3.2, we show that the model exhibits clustering, i.e., two vertices are more likely to be adjacent if they have a common neighbor. In social networks clustering translates into the idea that two people who share a common friend are more likely to know each other than people who do not. Thirdly in Section 2.3.3, we show that the Dense Random Dot Product Graph has a low fixed diameter of at most six. Finally, we end the chapter by extending the Dense Random Dot Product Graph into higher dimensions. We prove some results similar to those in a single dimension and discuss a bend that occurs in the degree distribution power law.

Chapter 3 discusses the Sparse Random Dot Product Graph. In Section 3.1.1, the sparse version of the model is introduced and simple results are proved. The main results parallel those in the dense case and are presented in Section 3.2. We show that under certain restrictions, the sparse model obeys the degree power law, exhibits clustering, and has a small diameter. Additionally, unlike in the dense model where all small subgraphs appear with probability tending to one, in the sparse model subgraphs appear at certain thresholds dependent upon a parameter b that is not present in the dense version. We present specific results regarding the thresholds for the appearance of edges, cliques, cycles, and trees and discuss the evolution of the Sparse Random Dot Product Graph as b goes from zero to infinity. Then in Section 3.2.1, we prove a general threshold result for the appearance of any graph H.

In Chapter 4 we briefly step away from the Random Dot Product Graphs to build a framework for the discussion of the discrete version of the model. A class of posynomials is introduced and an equivalence relation developed. Geometric techniques for studying the asymptotics of these posynomials is presented. These techniques will be used in Chapter 5 in first and second moment calculations.

Chapter 5 discusses the Discrete Random Dot Product Graph. In Section 5.1.1, the discrete version of the model is presented and a few basic results are proved. Unlike in the basic dense and sparse versions, we do not only consider the dimension one case, but instead draw the vectors from dimension $t \geq 1$. In Section 5.1.2, we illustrate the calculation for the threshold for the appearance of K_3 as a subgraph. In Section 5.2, we define the *probability order polynomial*, or POP, of a graph H as a function that is asymptotic to $P_{\geq}[H]$, the probability of H appearing as a (not necessarily induced) subgraph of a Discrete Random Dot Product Graph G. We give a general method for calculating the POP of H and present formulas for the POPs of trees, cycles, and complete graphs. In Section 5.3, using the framework built in Chapter 4, we present first moment results for trees, cycles, and complete graphs. We also prove a complete threshold result for K_3 and describe a general method for proving threshold results when all the required POPs are known.

Finally, in Chapter 6 we summarize the main results of this thesis and discuss avenues for additional research.

Chapter 2

The Dense Model

2.1 The Random Dot Product Graph – A New Model

In this thesis, we develop a set of models based on the notion that common interests affect relationships. Additionally, we reproduce some of the leading characteristics of real world social networks. The basic idea is as follows: Let G be graph on n vertices $V(G)$. With each vertex $v \in V(G)$ we associate an interest vector x_v and the probability that vertices u and v are adjacent is dependent on $x_u \cdot x_v$.

2.1.1 The General Model

Consider a graph G on n vertices $V(G)$. Let t be a positive integer and let $\mathrm{X} : V(G) \to \mathbb{R}^t$ be a mapping that assigns to each vertex $v \in V(G)$ a vector $\mathrm{X}(v) = x_v$.

Also, let $f : \mathbb{R} \to [0,1]$ be a function that maps the dot products of the vectors into probabilities. We define the Random Dot Product Graph G as follows:

- G has n vertices in $V(G)$,
- $\forall\, u, v \in V(G)$ the edge from u to v appears in G with probability $P_{\mathrm{X}}[u \sim v] = f(x_u \cdot x_v)$.

Let $\mathcal{G}_n = \{\text{all graphs on } V = \{1, \ldots, n\}\}$. We define the probability space $(\mathcal{G}_n, P_{\mathrm{X}})$ as follows: for any $H \in \mathcal{G}_n$

$$P_{\mathrm{X}}[H] = \left[\prod_{\substack{uv \in E(H)\\ u<v}} f(x_u \cdot x_v)\right]\left[\prod_{\substack{uv \notin E(H)\\ u<v}} (1 - f(x_u \cdot x_v))\right].$$

In the above discussion, we assume that for each vertex $v \in V(G)$ we are given the vectors x_v ahead of time, and based on these vectors some probability function f is chosen. However, we wish to model the social networks discussed above for which no such X is yet known. So, suppose that the vectors x_v are drawn independently from some random distribution and that the graph G is then generated using an appropriate choice of f. Then for any graph $H \in \mathcal{G}_n$

$$P[H] = \int P_{\mathrm{X}}[H]\, d\mathrm{X}.$$

In this thesis, we study the behavior of the random dot product graph when the x_v's are drawn from various distributions.

2.2 The Dense Random Dot Product Graph

In the rest of this chapter, we present and examine the Dense Random Dot Product Graph, the first of three versions of the Random Dot Product Graph discussed in this thesis. We begin by defining the model and discussing basic results. In Section 2.3.1, we show that the model obeys the degree distribution power law. In Section 2.3.2, we show that the model exhibits clustering. In Section 2.3.3, we show that the Dense Random Dot Product Graph has a low fixed diameter of at most six. Finally in Section 2.4, we end the chapter by extending the model into higher dimensions, proving basic results and discussing the appearance of a bend the degree distribution power law.

2.2.1 Some Results in the Simplest Case

We begin by looking at a relatively simple case. For each $v \in V(G)$ let the vector x_v be a one-dimensional vector drawn from $\mathcal{U}^a[0,1]$, the a^{th} power ($a > 1$) of the uniform distribution on $[0,1]$. Then, for any v,

$$P[x_v \leq r] = P[u^a \leq r] = P[u \leq r^{\frac{1}{a}}] = r^{\frac{1}{a}}$$

where $r \in [0,1]$ and $u \sim \mathcal{U}[0,1]$. Therefore $\forall\, [i,j] \subseteq [0,1]$,

$$P[x_v \in [i,j]] = \int_i^j \frac{1}{a} x^{\frac{1-a}{a}}\, dx$$

and the density function for x_v is $g(x) = \frac{1}{a}\, x^{\frac{1-a}{a}}$.

Now, in order to study the random dot product graph G we need to select a probability mapping f. The simplest choice of f is the identity function, $f(r) = r$. Note in the current setting, with the vectors drawn from $\mathcal{U}^a[0,1]$, the identity function does indeed map the dot products to values in $[0,1]$ and therefore into probabilities. Furthermore, we interpret the vectors as a level of interest in a given topic; i.e., 0 corresponds to no interest in the topic, and 1 a very high level of interest. Here we choose f so that if $x_v \cdot x_u$ is near 1, indicating that both u and v have a great interest in the topic, then the probability $f(x_v \cdot x_u)$ that they know each other is also near 1. Likewise, if $x_v \cdot x_u$ is small, we would like $f(x_v \cdot x_u)$ to be small. The identity function achieves this. So, from now on in this chapter, we assume that f is the identity function.

We denote this sample space of dense random dot product graphs on n vertices in which one-dimensional interest vectors are drawn from $\mathcal{U}^a[0,1]$ as $\mathcal{D}[n,a,1]$. Also, for any $u, v \in V(G)$, $u \sim v$ is defined as u is adjacent to v. We have the following results.

Proposition 2.2.1. *Let G be drawn from $\mathcal{D}[n,a,1]$. For any $u, v \in V(G)$ we have $P[u \sim v] = \frac{1}{(a+1)^2}$.*

Proof:

$$P[u \sim v] = \int_0^1 \int_0^1 x_u\, x_v\, g(x_u)\, g(x_v)\, dx_u\, dx_v = \frac{1}{a^2} \int_0^1 \int_0^1 x_u^{1/a}\, x_v^{1/a}\, dx_u\, dx_v = \frac{1}{(a+1)^2}.$$

QED.

So, an arbitrary edge appears in the graph with probability $\frac{1}{(a+1)^2}$ and the expected number of edges is $\binom{n}{2} \frac{1}{(a+1)^2}$.

Next, assume that n is large. We wish to study the degree distribution of G.

Proposition 2.2.2. *Let G be drawn from $\mathcal{D}[n, a, 1]$. The expected number of vertices of degree zero in G is*

$$E[|\{v : d(v) = 0\}|] \sim \left(\frac{1}{a} (1+a)^{\frac{1}{a}}\, \Gamma \left(\frac{1}{a} \right) \right) n^{\frac{a-1}{a}}.$$

Here, Γ is the classical gamma function $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$.

Proof:

Choose $v \in V(G)$ fixed, but arbitrary, and denote the vector of v by y. Let $x_1, x_2, \ldots, x_{n-1}$ be the vectors of the remaining $n-1$ vertices in $V(G) - \{v\}$. Then

$$P[d(v) = 0] = \int_0^1 \int_0^1 \cdots \int_0^1 (1 - x_1 y) \cdots (1 - x_{n-1} y)$$

$$\cdot g(x_1) \cdots g(x_{n-1})\, g(y)\, dx_1 \cdots dx_{n-1}\, dy$$

which is separable. Noting that for each of the x_i's

$$\int_0^1 (1 - x_i y) g(x_i) dx_i = \int_0^1 \frac{1}{a} (1 - x_i y) x_i^{\frac{1-a}{a}} \, dx_i = 1 - \frac{y}{a+1}$$

we have that

$$P[d(v) = 0] = \int_0^1 \left(1 - \frac{y}{a+1}\right)^{n-1} \frac{1}{a} y^{\frac{1-a}{a}} \, dy.$$

Substituting $\hat{y} = y^{1/a}$ for y in the integral we have that

$$P[d(v) = 0] = \int_0^1 \left(1 - \frac{\hat{y}^a}{a+1}\right)^{n-1} d\hat{y}.$$

From $\hat{y} = \frac{t}{n^{1/a}}$, one has $d\hat{y} = \frac{dt}{n^{1/a}}$ and $0 \leq \hat{y} \leq 1$ yields $0 \leq t \leq n^{1/a}$. Therefore, substituting again we have that

$$P[d(v) = 0] = \int_0^{n^{1/a}} \left(1 - \frac{t^a}{(a+1)n}\right)^{n-1} \frac{dt}{n^{1/a}}$$

$$\sim \int_0^{\infty} \exp\left\{\frac{-t^a}{a+1}\right\} \frac{dt}{n^{1/a}} = \left(\frac{1}{a} (1+a)^{\frac{1}{a}} \Gamma\left(\frac{1}{a}\right)\right) \frac{1}{n^{1/a}}.$$

Therefore the expected number of vertices of degree zero is

$$E[|\{v : d(v) = 0\}|] = n \, P[d(v) = 0] \sim \left(\frac{1}{a} (1+a)^{\frac{1}{a}} \Gamma\left(\frac{1}{a}\right)\right) n^{\frac{a-1}{a}}.$$

QED.

We see that the expected number of isolated vertices in G is a constant times

$n^{\frac{a-1}{a}}$. The above result can be generalized to vertices of degree k where $k \ll n$ as follows:

Proposition 2.2.3. *Let k be a fixed nonnegative integer. Let $\lambda(k)$ be the number of vertices of degree k in a random dot product graph drawn from $\mathcal{D}[n, a, 1]$. Then*

$$E[\lambda(k)] \sim \left(\frac{1}{k!\,a}(1+a)^{\frac{1}{a}}\Gamma\left(\frac{1}{a}+k\right)\right) n^{\frac{a-1}{a}}$$

as $n \to \infty$.

2.3 Main Results

2.3.1 Degree Power Law

Now, we wish to show that the Dense Random Dot Product Graph is a good candidate for modeling social networks, including the internet. To this end, we posit the following.

Conjecture 2.3.1. *Let $\varepsilon > 0$ and let n, k be integers with $\varepsilon n < k < (1-\varepsilon)n$. Let $\lambda(k)$ be the number of vertices of degree k in a graph drawn from $\mathcal{D}[n, a, 1]$. Then as $n \to \infty$, with high probability $\lambda(k)$ satisfies the degree distribution power law $\lambda(k) \propto k^{\phi}$, with $\phi = \frac{1-a}{a}$.*

We observe that a log-log histogram of the degree distribution is a straight line over most of the degree sequence of Dense Random Dot Product Graphs. While

Conjecture 2.3.1 would explain this directly, a mild relaxation of this conjecture also explains this phenomenon. For positive integers a, b with $a < b$, define $\lambda[a, b]$ to be the number of vertices in a graph with degrees in the interval $[a, b]$. We show that $\lambda[k(1-\delta), k(1+\delta)]$ is proportional to $(2k\delta)k^{\phi}$ for most values of k and for almost all Dense Random Dot Product Graphs (i.e., with high probability as $n \to \infty$). The basic idea of the proof is as follows.

For a given k and δ, we want to know $\lambda[k(1-\delta), k(1+\delta)]$. However, we do not count the number of vertices whose degrees fall in $[k(1-\delta), k(1+\delta)]$ directly. Instead we select a value $s \in [0, 1]$ so that if $x_v = s$ then $E[d(v)|x_v = s] = \frac{(n-1)x_v}{a+1} = \frac{(n-1)s}{a+1} = k$ (note that this establishes a direct relationship between s and k). Then we count the number of vertices whose vectors fall in the interval $\mathcal{S} = [s(1-\delta), s(1+\delta)]$, since for any vertex v with $x_v \in \mathcal{S}$ the $E[d(v)|x_v] = \frac{(n-1)x_v}{a+1} \in \left[\frac{(n-1)s(1-\delta)}{a+1}, \frac{(n-1)s(1+\delta)}{a+1}\right] = [k(1-\delta), k(1+\delta)]$. Likewise, if $x_v \notin \mathcal{S}$ then $E[d(v)|x_v] \notin [k(1-\delta), k(1+\delta)]$. So, the number of vertices with $x_v \in \mathcal{S}$ is expected to be the same as the number of vertices with $d(v) \in [k(1-\delta), k(1+\delta)]$.

Now, to allow for the variance that occurs in our model, we look at intervals $\mathcal{S}_-$ and $\mathcal{S}_+$ that are slightly smaller and larger than $\mathcal{S}$, respectively. We show that $\{v : x_v \in \mathcal{S}_-\} \subseteq \{v : d(v) \in [k(1-\delta), k(1+\delta)]\} \subseteq \{v : x_v \in \mathcal{S}_+\}$, each containment occurring with high probability. And therefore, the number of vertices whose degrees fall in $[k(1-\delta), k(1+\delta)]$ must be bounded by the number of vertices whose vectors fall in $\mathcal{S}_-$ and $\mathcal{S}_+$.

Before we begin, we define an $\mathbf{X}$-labeled Dense Random Dot Product Graph. As with an ordinary Dense Random Dot Product Graph, for each $v \in V(G)$, let the vector x_v be a one-dimensional vector drawn from $\mathcal{U}^a[0,1]$ and let the probability mapping f be the identity function. Additionally, let $\mathbf{X}$ be the $1 \times n$ matrix of vectors. We denote this sample space of $\mathbf{X}$-labeled Dense Random Dot Product Graphs on n vertices as $\mathcal{D}[\mathbf{X}]$. Note that the only difference between $G \in \mathcal{D}[n, a, 1]$ and $(G, \mathbf{X}) \in \mathcal{D}[\mathbf{X}]$ is that when G is a $\mathbf{X}$-labeled Dense Random Dot Product Graph the vectors are retained.

Lemma 2.3.2. *Let $(G, \mathbf{X})$ be drawn from $\mathcal{D}[\mathbf{X}]$. Let $s \in (n^{-1/24}, 1)$, $\delta = n^{-1/12}$, and $\varepsilon = n^{-1/3}$. Define the interval $\mathcal{S}_- = [\frac{s(1-\delta)}{1-\varepsilon}, \frac{s(1+\delta)}{1+\varepsilon}]$. Then*

$$\mu_{\mathcal{S}_-} = E[d(v)|x_v \in \mathcal{S}_-] \in \left[\frac{s(1-\delta)(n-1)}{(1-\varepsilon)(a+1)}, \frac{s(1+\delta)(n-1)}{(1+\varepsilon)(a+1)}\right].$$

Proof:

$$\begin{aligned}
\mu_{\mathcal{S}_-} = E[d(v)|x_v \in \mathcal{S}_-] &= E\left[\sum_{w \in V(G)\,:\,w \neq v} \mathbf{I}\{v \sim w | x_v \in \mathcal{S}_-\}\right] \\
&= (n-1)P[v \sim w | x_v \in \mathcal{S}_-] \\
&= (n-1)\frac{\int_{\frac{s(1-\delta)}{1-\varepsilon}}^{\frac{s(1+\delta)}{1+\varepsilon}} \int_0^1 x_v x_u g(x_v) g(x_u) dx_v dx_u}{\int_{\frac{s(1-\delta)}{1-\varepsilon}}^{\frac{s(1+\delta)}{1+\varepsilon}} g(x_v) dx_v} \\
&= (n-1)\frac{\left(\int_{\frac{s(1-\delta)}{1-\varepsilon}}^{\frac{s(1+\delta)}{1+\varepsilon}} x_v g(x_v) dx_v\right)\left(\int_0^1 x_u g(x_u) dx_u\right)}{\int_{\frac{s(1-\delta)}{1-\varepsilon}}^{\frac{s(1+\delta)}{1+\varepsilon}} g(x_v) dx_v}.
\end{aligned}$$

We look at each of the integrals separately. First note that $\int_0^1 x_u g(x_u)dx_u = \frac{1}{a+1}$. Next we note that

$$\int_{\frac{s(1-\delta)}{1-\varepsilon}}^{\frac{s(1+\delta)}{1+\varepsilon}} g(x_v)dx_v = \left[\left(\frac{s(1+\delta)}{1+\varepsilon}\right)^{\frac{1}{a}} - \left(\frac{s(1-\delta)}{1-\varepsilon}\right)^{\frac{1}{a}}\right] = \left[\hat{\delta}\frac{1}{a}s^{*\frac{1}{a}-1}\right]$$

for $\hat{\delta} = \left(\frac{s(1+\delta)}{1+\varepsilon}\right) - \left(\frac{s(1-\delta)}{1-\varepsilon}\right)$ and some $s^* \in \left[\frac{s(1-\delta)}{1-\varepsilon}, \frac{s(1+\delta)}{1+\varepsilon}\right]$ by the mean value theorem.

Similarly

$$\int_{\frac{s(1-\delta)}{1-\varepsilon}}^{\frac{s(1+\delta)}{1+\varepsilon}} x_v g(x_v)dx_v = \frac{1}{1+a}\left[\hat{\delta}\frac{a+1}{a}s^{**\frac{1}{a}}\right]$$

for some $s^{**} \in \left[\frac{s(1-\delta)}{1-\varepsilon}, \frac{s(1+\delta)}{1+\varepsilon}\right]$.

So we have

$$\mu_{\mathcal{S}_-} = \frac{(n-1)}{(a+1)}\frac{\frac{1}{a+1}\left[\hat{\delta}\frac{a+1}{a}s^{**\frac{1}{a}}\right]}{\left[\hat{\delta}\frac{1}{a}s^{*\frac{1}{a}-1}\right]} = \frac{(n-1)}{(a+1)}s^{*1-\frac{1}{a}}s^{**\frac{1}{a}}.$$

Now $\left(\frac{s(1-\delta)}{1-\varepsilon}\right) \leq s^*, s^{**} \leq \left(\frac{s(1+\delta)}{1+\varepsilon}\right)$. Therefore $\left(\frac{s(1-\delta)}{1-\varepsilon}\right) \leq s^{*1-\frac{1}{a}}s^{**\frac{1}{a}} \leq \left(\frac{s(1+\delta)}{1+\varepsilon}\right)$. So, we have

$$\frac{s(1-\delta)}{1-\varepsilon}\frac{n-1}{a+1} \leq \mu_{\mathcal{S}_-} \leq \frac{s(1+\delta)}{1+\varepsilon}\frac{n-1}{a+1}.$$

QED.

We have bounded the expected degree of any vertex whose vector falls in $\mathcal{S}_-$.

Lemma 2.3.3. *Let $(G, \mathbf{X})$ be drawn from $\mathcal{D}[\mathbf{X}]$. Let $s, \mathcal{S}_-, \mu_{\mathcal{S}_-}, \delta$ and ε be defined as in Lemma 2.3.2. Let $v \in V(G)$. If $\frac{s(1-\delta)}{1-\varepsilon} \leq x_v \leq \frac{s(1+\delta)}{1+\varepsilon}$, then with probability tending* *to* 1 *as* $n \to \infty$

$$d(v) \in \left[\frac{s(1-\delta)(n-1)}{a+1}, \frac{s(1+\delta)(n-1)}{a+1}\right].$$

Indeed the probability $d(v) \notin \left[\frac{s(1-\delta)(n-1)}{a+1}, \frac{s(1+\delta)(n-1)}{a+1}\right]$ *goes to zero faster than the reciprocal of any polynomial in* n.

Proof:

Let $v \in V(G)$ and $\frac{s(1-\delta)}{1-\varepsilon} \leq x_v \leq \frac{s(1+\delta)}{1+\varepsilon}$. Now, the degree of a vertex $d(v)$ is the sum over w of the iid indicator variables $\mathbf{I}\{v \sim w\}$ and so by Chernoff's bounds we have $P[d(v) < (1-\varepsilon)\mu_{\mathcal{S}_-}] \leq \exp\left\{\frac{-\varepsilon^2 \mu_{\mathcal{S}_-}}{3}\right\}$. And so by Lemma 2.3.2 we have

$$\begin{aligned} P\left[d(v) < (1-\delta)s\frac{n-1}{a+1}\right] &= P\left[d(v) < (1-\varepsilon)\frac{(1-\delta)}{1-\varepsilon}s\frac{n-1}{a+1}\right] \\ &\leq P[d(v) < (1-\varepsilon)\mu_{\mathcal{S}_-}] \\ &\leq \exp\left\{\frac{-\varepsilon^2 \mu_{\mathcal{S}_-}}{3}\right\} \\ &\leq \exp\left\{\frac{-\varepsilon^2(1-\delta)(n-1)s}{(1-\varepsilon)(a+1)3}\right\} \\ &\leq \exp\left\{\frac{-\varepsilon^2(1-\delta)(n-1)s}{(a+1)3}\right\} \\ &\leq \exp\left\{\frac{-n^{-2/3}(1-n^{-1/12})(n-1)n^{-1/24}}{(a+1)3}\right\} \to 0 \end{aligned}$$

as $n \to \infty$.

Similarly

$$P\left[d(v) > (1+\delta)s\frac{n-1}{a+1}\right] \leq \exp\left\{\frac{-\varepsilon^2(1-\delta)(n-1)s}{(a+1)3}\right\}$$

$$\leq \exp\left\{\frac{-n^{-2/3}(1-n^{-1/12})(n-1)n^{-1/24}}{(a+1)3}\right\} \to 0$$

as $n \to \infty$. And so with probability tending to 1 as $n \to \infty$

$$d(v) \in \left[\frac{s(1-\delta)(n-1)}{a+1}, \frac{s(1+\delta)(n-1)}{a+1}\right].$$

QED.

And so we have shown that with high probability if $x_v \in \mathcal{S}_-$ then $d(v) \in [k(1-\delta), k(1+\delta)]$ since $k = \frac{(n-1)s}{a+1}$.

Lemma 2.3.4. *Let $(G, \mathbf{X})$ be drawn from $\mathcal{D}[\mathbf{X}]$ and $k \geq n^{23/24}$. Let $s = \frac{k(n-1)}{a+1}$ and define the interval $\mathcal{S}_+ = [(1-n^{-1/3})s(1-n^{-1/12}), (1+n^{-1/3})s(1+n^{-1/12})]$. Let $v \in V(G)$. If $k(1-n^{-1/12}) \leq d(v) \leq k(1+n^{-1/12})$ then with probability tending to 1 as $n \to \infty$*

$$x_v \in \mathcal{S}_+ = [(1-n^{-1/3})s(1-n^{-1/12}), (1+n^{-1/3})s(1+n^{-1/12})].$$

Indeed the probability $x_v \notin \mathcal{S}_+$ goes to zero faster than a reciprocal of any polynomial in n.

Proof:

Let $d(v) \in [k(1-n^{-1/12}), k(1+n^{-1/12})]$. First we examine what occurs when the vectors fall below the interval $\mathcal{S}_+$. By way of contradiction, assume $x_v < (1-n^{-1/3})s(1-n^{-1/12})$.

Consider the case when $n^{-1/4} \leq x_v < (1-n^{-1/3})s(1-n^{-1/12})$. In this case the Chernoff bound gives us that

$$P[d(v) > (1+n^{-1/3})E[d(v)|x_v]] \leq \exp\left\{\frac{-n^{-2/3}E[d(v)|x_v]}{3}\right\}$$

$$= \exp\left\{\frac{-n^{-2/3}(n-1)x_v}{3(a+1)}\right\}$$

$$\leq \exp\left\{\frac{-(n^{1/3}-n^{-2/3})n^{-1/4}}{3(a+1)}\right\}$$

since $x_v \geq n^{-1/4}$.

Also, $x_v < (1-n^{-1/3})s(1-n^{-1/12})$ gives us that

$$P[d(v) > (1+n^{-1/3})E[d(v)|x_v]] = P\left[d(v) > (1+n^{-1/3})\frac{(n-1)x_v}{a+1}\right]$$

$$\geq P[d(v) > (1+n^{-1/3})\frac{(n-1)}{(a+1)}(1-n^{-1/3})s(1-n^{-1/12})]$$

$$= P[d(v) > k(1+n^{-1/3})(1-n^{-1/3})(1-n^{-1/12})]$$

$$= P[d(v) > k(1-n^{-1/12})(1-n^{-2/3})]$$

$$= P[d(v) > k(1 - n^{-1/12}) + k(1 - n^{-1/12})(-n^{-2/3})]$$

$$= P[d(v) > k(1 - n^{-1/12}) + k(n^{-3/4} - n^{-2/3})]$$

$$\geq P[d(v) > k(1 - n^{-1/12}) + n^{23/24}(n^{-3/4} - n^{-2/3})]$$

(since $k \geq n^{23/24}$ and $n^{-3/4} - n^{-2/3} < 0$ for large n)

$$= P[d(v) > k(1 - n^{-1/12}) + (n^{5/24} - n^{7/24})]$$

$$\geq P[d(v) > k(1 - n^{-1/12}) + 1]$$

(since $n^{5/24} - n^{7/24} \leq 1$ for large n)

$$= P[d(v) \geq k(1 - n^{-1/12})].$$

Therefore

$$P[d(v) \geq k(1 - n^{-1/12})] \leq \exp\left\{\frac{-(n^{1/3} - n^{-2/3})n^{-1/4}}{3(a+1)}\right\}.$$

So,

$$E[\,|\{v : d(v) \geq k(1 - n^{-1/12})\}|\,|n^{-1/4} \leq x_v < (1 - n^{-1/3})s(1 - n^{-1/12})]$$

$$\leq n\, P[d(v) \geq k(1 - n^{-1/12})\big|n^{-1/4} \leq x_v < (1 - n^{-1/3})s(1 - n^{-1/12})]$$

$$\leq n \exp\left\{ \frac{-(n^{1/3} - n^{-2/3})n^{-1/4}}{3(a+1)} \right\}$$

$$\leq n \exp\left\{ \frac{-(n^{1/12} - n^{-11/12})}{3(a+1)} \right\} \to 0$$

as $n \to \infty$, since $\log n - (n^{1/12} - n^{-11/12}) \to -\infty$. Therefore by Markov's inequality

$$P[\,|\{v : d(v) \geq k(1 - n^{-1/12})\}| = 0 \;|n^{-1/4} \leq x_v < (1 - n^{-1/3})s(1 - n^{-1/12})] \to 1.$$

So, if $n^{-1/4} \leq x_v < (1 - n^{-1/3})s(1 - n^{-1/12})$ then, with probability tending to 1, $d(v) < k(1 - n^{-1/12})$ which is a contradiction.

Now consider the case when $0 \leq x_v \leq n^{-1/4}$. In this case

$$E[d(v)|0 \leq x_v \leq n^{-1/4}] = E\left[\sum_{w \in V(G)} \mathbf{I}\left\{ v \sim w | 0 \leq x_v \leq n^{-1/4} \right\} \right]$$

$$= (n-1)P[v \sim w | 0 \leq x_v \leq n^{-1/4}]$$

$$= (n-1) \frac{\int_0^{n^{-1/4}} \int_0^1 x_v x_w f(x_v) f(x_w) dx_v dx_w}{\int_0^{n^{-1/4}} f(x_v) dx_v}$$

$$= \frac{(n-1)n^{-1/4}}{(a+1)^2}.$$

So the above equation and the fact that $k \geq n^{23/24}$ give us that

$$P[d(v) > k(1 - n^{-1/12})] \leq P[d(v) > n^{23/24}(1 - n^{-1/12})]$$

$$= P[d(v) > n^{23/24} - n^{21/24}]$$

$$\leq P\left[d(v) > (1 + n^{-1/3})\left(\frac{(n-1)n^{-1/4}}{(a+1)^2}\right)\right]$$

(since $n^{23/24} - n^{21/24} > (1 + n^{-1/3})\left(\frac{(n-1)n^{-1/4}}{(a+1)^2}\right)$)

$$= P[d(v) > (1 + n^{-1/3})E[d(v)|0 \leq x_v \leq n^{-1/4}]].$$

And we see from the Chernoff bound that

$$P[d(v) > k(1 - n^{-1/12})] \leq \exp\left\{\frac{(-n^{-2/3})E[d(v)|0 \leq x_v \leq n^{-1/4}]}{3}\right\}$$

$$\exp\left\{\frac{(-n^{-2/3})(n-1)n^{-1/4}}{3(a+1)^2}\right\}.$$

So,

$$E[\,|\{v : d(v) \geq k(1 - n^{-1/12})\}|\,|0 \leq x_v \leq n^{-1/4}]$$

$$\leq n\,P[d(v) \geq k(1 - n^{-1/12})\,|0 \leq x_v \leq n^{-1/4}]$$

$$\leq n \exp\left\{\frac{(-n^{-2/3})(n-1)n^{-1/4}}{3(a+1)^2}\right\}$$

$$\leq n \exp\left\{\frac{-(n^{1/12} - n^{-11/12})}{3(a+1)^2}\right\} \to 0$$

as $n \to \infty$, since $\log n - (n^{1/12} - n^{-11/12}) \to -\infty$. Therefore by Markov's inequality $P[\,|\{v : d(v) \geq k(1 - n^{-1/12})\}| = 0] \to 1$. So, if $0 \leq x_v \leq n^{-1/4}$ then, with probability tending to 1, $d(v) < k(1 - n^{-1/12})$ which is a again a contradiction.

Therefore, since both cases create a contradiction, with high probability $x_v \geq (1 - n^{-1/3})s(1 - n^{-1/12})$ and the vectors do not fall below the interval $\mathcal{S}_+$.

Now we look for a contradiction when the interest are above the interval $\mathcal{S}_+$. By way of contradiction, assume $x_v > (1+n^{-1/3})s(1+n^{-1/12})$. The Chernoff bound gives us

$$P[d(v) < (1 - n^{-5/12})E[d(v)|x_v] \leq \exp\left\{-n^{-5/6}\frac{(n-1)x_v}{(a+1)3}\right\}$$

$$\leq \exp\left\{-n^{-5/6}\frac{(n-1)}{(a+1)3}(1 + n^{-1/3})s(1 + n^{-1/12})\right\}$$

$$= \exp\left\{-n^{-5/6}\frac{k}{3}(1 + n^{-1/12} + n^{-1/3} + n^{-5/12})\right\}$$

(since $k \geq n^{23/24}$)

$$\leq \exp\left\{-n^{-5/6}\frac{n^{23/24}}{3}(1 + n^{-1/12} + n^{-1/3} + n^{-5/12})\right\}$$

$$= \exp\left\{(n^{1/8} + n^{1/24} + n^{-5/24} + n^{-7/12})/3\right\} \to 0$$

as $n \to \infty$.

Also,

$$P[d(v) < (1-n^{-5/12})\frac{(n-1)}{(a+1)}x_v] \geq P[d(v) < (1-n^{-5/12})\frac{(n-1)}{(a+1)}(1+n^{-1/3})s(1+n^{-1/12})]$$

$$= P[d(v) < k(1 - n^{-5/12})(1 + n^{-1/3})(1 + n^{-1/12})]$$

$$= P[d(v) < k(1 - n^{-1/12})(1 + n^{-5/12} + n^{-1/3} - n^{-3/4})]$$

$$= P[d(v) < k(1 + n^{-1/12}) + k(1 + n^{-1/12})(n^{-5/12} + n^{-1/3} - n^{-3/4})]$$

$$= P[d(v) < k(1 + n^{-1/12}) + k(n^{-1/3} - n^{-3/4} - n^{-1/2} - n^{-5/6})]$$

$$\geq P[d(v) < k(1 + n^{-1/12}) + n^{23/24}(n^{-1/3} - n^{-3/4} - n^{-1/2} - n^{-5/6})]$$

(since $k \geq n^{23/24}$ and $n^{-1/3} - n^{-3/4} - n^{-1/2} - n^{-5/6} > 0$ for large n)

$$= P[d(v) < k(1 + n^{-1/12}) + (n^{5/8} - n^{5/24} - n^{11/24} - n^{1/8})]$$

$$\geq P[d(v) < k(1 + n^{-1/12}) + 1]$$

(since $n^{5/8} - n^{5/24} - n^{11/24} - n^{1/8} > 1$ for large n)

$$= P[d(v) \leq k(1 + n^{-1/12})].$$

So, $P[d(v) \leq k(1 + n^{-1/12})] \to 0$ as $n \to \infty$.

So,

$$E[\,|\{v : d(v) \leq k(1 + n^{-1/12})\}|\,|x_v > (1 + n^{-1/3})s(1 + n^{-1/12})]$$

$$\leq n\,P[d(v) \leq k(1 + n^{-1/12})\}|\,|x_v > (1 + n^{-1/3})s(1 + n^{-1/12})]$$

$$\leq n \exp\left\{-(n^{1/8}+n^{1/24}+n^{-5/24}+n^{-7/12})/3\right\} \to 0$$

as $n \to \infty$, since $\log n - (n^{1/8}+n^{1/24}+n^{-5/24}+n^{-7/12}) \to -\infty$. Therefore by Markov's inequality $P[\,|\{v : d(v) \leq k(1+n^{-1/12})\}| = 0] \to 1$. So, if $x_v > (1+n^{-1/3})s(1+n^{-1/12})$ then, with probability tending to 1, $d(v) > k(1+n^{-1/12})$ which is a contradiction. Therefore, with probability tending to 1, $x_v \leq (1+n^{-1/3})s(1+n^{-1/12})$ as $n \to \infty$. QED.

And so we see that whenever a vertex has $d(v) \in [k(1-n^{-1/12}), \leq k(1+n^{-1/12})]$ then $x_v \in \mathcal{S}_+$ with probability tending to 1 as $n \to \infty$.

Theorem 2.3.5. *Let $(G, \mathbf{X})$ be drawn from $\mathcal{D}[\mathbf{X}]$and $k \geq n^{23/24}$. Let $s = \frac{k(n-1)}{a+1}, \delta = n^{-1/12}$ and $\varepsilon = n^{-1/3}$. Define $\mathcal{S}_- = [\frac{s(1-\delta)}{1-\varepsilon}, \frac{s(1+\delta)}{1+\varepsilon}]$ and $\mathcal{S}_+ = [(1-\varepsilon)s(1-\delta), (1+\varepsilon)s(1+\delta)]$. Then $\{v : x_v \in \mathcal{S}_-\} \subseteq \{v : d(v) \in [k(1-\delta), k(1+\delta)]\} \subseteq \{v : x_v \in \mathcal{S}_+\}$ and*

$$\lambda[k(1-\delta), k(1+\delta)] = \left(\frac{n^{1-\frac{1}{a}}(a+1)^{\frac{1}{a}}}{a}\right)(2\,k[(\delta+\mathcal{O}(\varepsilon))(1+\mathcal{O}(\delta))^{\frac{1}{a}-1})]k^{\frac{1}{a}-1}(1+\mathcal{O}(\varepsilon))$$

with probability $\to 1$ as $n \to \infty$.

Proof:

Take $v \in \{v : x_v \in \mathcal{S}_-\}$. Note that $k \geq n^{23/24}$ implies that $s \geq n^{-1/24}$ and we have satisfied all of the conditions of Lemma 2.3.3. Therefore with probability tending to 1 as $n \to \infty$ we have that $d(v) \in \left[\frac{s(1-\delta)(n-1)}{a+1}, \frac{s(1+\delta)(n-1)}{a+1}\right] = [k(1-\delta), k(1+\delta)]$.

Therefore $v \in \{v : d(v) \in [k(1-\delta), k(1+\delta)]\}$ and

$$P\left[\{v : x_v \in \mathcal{S}_-\} \subseteq \{v : d(v) \in [k(1-\delta), k(1+\delta)]\}\right] \to 1$$

as $n \to \infty$.

Similarly, take $v \in \{v : d(v) \in [k(1-\delta), k(1+\delta)]\}$. Here all of the conditions of Lemma 2.3.4 are satisfied. Therefore with probability tending to 1 as $n \to \infty$ we have that $x_v \in \mathcal{S}_+$ and so $v \in \{v : x_v \in \mathcal{S}_+\}$. Hence

$$P\left[\{v : d(v) \in [k(1-\delta), k(1+\delta)]\} \subseteq \{v : x_v \in \mathcal{S}_+\}\right] \to 1$$

as $n \to \infty$.

We have shown that with high probability $\{v : x_v \in \mathcal{S}_-\} \subseteq \{v : d(v) \in [k(1-\delta), k(1+\delta)]\} \subseteq \{v : x_v \in \mathcal{S}_+\}$ and so $|\{v : x_v \in \mathcal{S}_-\}| \leq \lambda[k(1-\delta), k(1+\delta)] \leq |\{v : x_v \in \mathcal{S}_+\}|$.

Now

$$\begin{aligned} E[|\{v : x_v \in \mathcal{S}_-\}|] &= E\left[\sum_{v \in V(G)} \mathbf{I}\{x_v \in \mathcal{S}_-\}\right] \\ &= nP[x_v \in \mathcal{S}_-] \\ &= n \int_{\frac{s(1-\delta)}{1-\varepsilon}}^{\frac{s(1+\delta)}{1+\varepsilon}} g(x_v) dx_v \end{aligned}$$

which we know from the proof of Lemma 2.3.2 to be

$$n\left[\hat{\delta}\frac{1}{a}s^{*\frac{1}{a}-1}\right]$$

for $\hat{\delta} = \left(\frac{s(1+\delta)}{1+\varepsilon}\right) - \left(\frac{s(1-\delta)}{1-\varepsilon}\right)$ and some $s^* \in [\left(\frac{s(1-\delta)}{1-\varepsilon}\right), \left(\frac{s(1+\delta)}{1+\varepsilon}\right)]$.

Now, $\frac{1}{1-\varepsilon} = 1 + \varepsilon + \frac{\varepsilon^2}{1-\varepsilon}$. So,

$$\frac{s(1-\delta)}{1-\varepsilon} = s(1-\delta)(1+\varepsilon+\frac{\varepsilon^2}{1-\varepsilon}) = s - s(\delta - \varepsilon + \delta\varepsilon + \frac{\delta\varepsilon^2}{1-\varepsilon} - \frac{\varepsilon^2}{1-\varepsilon})$$

$$= s - s\delta - s\,\mathcal{O}(\varepsilon).$$

Likewise $\frac{1}{1+\varepsilon} = 1 - \varepsilon + \frac{\varepsilon^2}{1+\varepsilon}$ and

$$\frac{s(1+\delta)}{1+\varepsilon} = s + s\delta - s\,\mathcal{O}(\varepsilon).$$

And so we know that $\hat{\delta} = 2s\delta - s\,\mathcal{O}(\varepsilon)$. Also since $s, s^* \in [\left(\frac{s(1-\delta)}{1-\varepsilon}\right), \left(\frac{s(1+\delta)}{1+\varepsilon}\right)]$, we have that $s^* = s + s\,\mathcal{O}(\delta)$.

So we can replace s^* giving us that

$$E[|\{v : x_v \in \mathcal{S}_-\}|] = n\left[\hat{\delta}\frac{1}{a}s^{*\frac{1}{a}-1}\right] = n\left[(2s\delta - s\,\mathcal{O}(\varepsilon))\frac{1}{a}(s + s\mathcal{O}(\delta))^{\frac{1}{a}-1}\right]$$

$$= n\left[(2\delta - \mathcal{O}(\varepsilon))\frac{1}{a}s^{\frac{1}{a}}(1+\mathcal{O}(\delta))^{\frac{1}{a}-1}\right]$$

and substituting $s = \frac{k(a-1)}{n-1}$ we have

$$= n\left[(2\delta - \mathcal{O}(\varepsilon))\frac{1}{a}\left(\frac{k(a-1)}{n-1}\right)^{\frac{1}{a}}(1+\mathcal{O}(\delta))^{\frac{1}{a}-1}\right]$$

$$= \left(\frac{n^{1-\frac{1}{a}}(a+1)^{\frac{1}{a}}}{a}\right)(2\,k[(\delta - \mathcal{O}(\varepsilon))(1+\mathcal{O}(\delta))^{\frac{1}{a}-1})]k^{\frac{1}{a}-1}.$$

For any $0 < \hat{\varepsilon} < 1$ the Chernoff bounds give us that

$$P[|\{v : x_v \in \mathcal{S}_-\}| < (1-\hat{\varepsilon})E[|\{v : x_v \in \mathcal{S}_-\}|] \leq \exp\left\{\frac{-\hat{\varepsilon}^2 E[|\{v : x_v \in \mathcal{S}_-\}|]}{3}\right\}$$

$$= \exp\left\{\frac{-\hat{\varepsilon}^2\left(\frac{n^{1-\frac{1}{a}}(a+1)^{\frac{1}{a}}}{a}\right)(2\,k[(\delta - \mathcal{O}(\varepsilon))(1+\mathcal{O}(\delta))^{\frac{1}{a}-1})]k^{\frac{1}{a}-1}}{3}\right\}$$

$$= \exp\left\{-\hat{\varepsilon}^2\left(\frac{n^{1-\frac{1}{a}}(a+1)^{\frac{1}{a}}}{3a}\right)(2\,k[(n^{-1/12} - \mathcal{O}(n^{-1/3}))(1+\mathcal{O}(n^{-1/12}))^{\frac{1}{a}-1})]k^{\frac{1}{a}-1}\right\}$$

$$= \exp\left\{-\hat{\varepsilon}^2\left(\frac{n^{-1/a}(a+1)^{\frac{1}{a}}}{3a}\right)2\,[(n^{11/12} - \mathcal{O}(n^{2/3}))(1+\mathcal{O}(n^{-1/12}))^{\frac{1}{a}-1})]k^{\frac{1}{a}}\right\}$$

$$= \exp\left\{-\hat{\varepsilon}^2\left(\frac{n^{-1/a}(a+1)^{\frac{1}{a}}}{3a}\right)2\,[(n^{11/12} - \mathcal{O}(n^{2/3}))(1+\mathcal{O}(n^{-1/12}))^{\frac{1}{a}-1})](n^{23/24a})\right\}$$

since $k \geq n^{23/24}$. If we let $\hat{\varepsilon} = \varepsilon = n^{-1/3}$ we have

$$P[|\{v : x_v \in \mathcal{S}_-\}| < (1-n^{-1/3})E[|\{v : x_v \in \mathcal{S}_-\}|]$$

$$= \exp\left\{-n^{-2/3}\left(\frac{n^{-1/a}(a+1)^{\frac{1}{a}}}{3a}\right)2\,[(n^{11/12} - \mathcal{O}(n^{2/3}))(1+\mathcal{O}(n^{-1/12}))^{\frac{1}{a}-1})](n^{23/24a})\right\}$$

$$= \exp\left\{-\left(\frac{n^{-1/a}(a+1)^{\frac{1}{a}}}{3a}\right)2\left[(n^{3/12}-\mathcal{O}(1))(1+\mathcal{O}(n^{-1/12}))^{\frac{1}{a}-1})\right](n^{23/24a})\right\} \to 0$$

as $n \to \infty$.

Similarly

$$E[|\{v : x_v \in \mathcal{S}_+\}|] = \left(\frac{n^{1-\frac{1}{a}}(a+1)^{\frac{1}{a}}}{a}\right) 2\,k[(\delta+\mathcal{O}(\varepsilon))(1+\mathcal{O}(\delta))^{\frac{1}{a}-1})]k^{\frac{1}{a}-1},$$

and likewise,

$$P[|\{v : x_v \in \mathcal{S}_+\}| > (1+n^{-1/3})E[|\{v : x_v \in \mathcal{S}_+\}|] \to 0$$

as $n \to \infty$.

So we have that

$$(1-\varepsilon)\left(\frac{n^{1-\frac{1}{a}}(a+1)^{\frac{1}{a}}}{a}\right)(2\,k[(\delta-\mathcal{O}(\varepsilon))(1+\mathcal{O}(\delta))^{\frac{1}{a}-1})]k^{\frac{1}{a}-1} \leq \lambda[k(1-\delta),k(1+\delta)]$$

$$\leq (1+\varepsilon)\left(\frac{n^{1-\frac{1}{a}}(a+1)^{\frac{1}{a}}}{a}\right)(2\,k[(\delta+\mathcal{O}(\varepsilon))(1+\mathcal{O}(\delta))^{\frac{1}{a}-1})]k^{\frac{1}{a}-1}$$

with probability tending to 1 as $n \to \infty$. And therefore

$$\lambda[k(1-\delta),k(1+\delta)] = \left(\frac{n^{1-\frac{1}{a}}(a+1)^{\frac{1}{a}}}{a}\right)(2\,k[(\delta+\mathcal{O}(\varepsilon))(1+\mathcal{O}(\delta))^{\frac{1}{a}-1})]k^{\frac{1}{a}-1}(1+\mathcal{O}(\varepsilon))$$

with probability $\to 1$ as $n \to \infty$.

QED.

2.3.2 Clustering

In an Erdős-Rényi random graph, the existence of a given edge is independent of the rest of the graph. In other words, the probability that v and w are adjacent is unaffected by the existence of any other edges. In particular, for any three distinct vertices u, v, w, we have $P(u \sim w | u \sim v \sim w) = P(u \sim w)$.

The situation in Dense Random Dot Product Graphs is different; these graphs more accurately reflect the communication networks they are intended to model. Intuitively, suppose u, v, w are agents, and we know that u communicates with v, and v communicates with w. Then this knowledge increases the likelihood that u and w communicate by virtue of their common acquaintance. That is, we would expect to see that $P(u \sim w | u \sim v \sim w) > P(u \sim w)$. We can derive exactly this result in the Dense Random Dot Product Graph model.

Lemma 2.3.6. *Let G be drawn from $\mathcal{D}[n, a, 1]$. Let $v \in V(G)$ and let $N(v)$ be the open neighborhood of v. Then $\forall\, u, w \in N(v)$,*

$$P[u \sim w | u \sim v \sim w] = \left(\frac{a+1}{2a+1}\right)^2.$$

Proof:

Let x_v, x_u, and x_w be the vectors of v, u and w, respectively. Then

$$P[u \sim w | u \sim v \sim w] = \frac{P[uvw \text{ is a triangle}]}{P[u \sim v \sim w]}$$

$$= \frac{\int_0^1 \int_0^1 \int_0^1 x_v^2 x_u^2 x_w^2 g(x_v) g(x_u) g(x_w) dx_v dx_u dx_w}{\int_0^1 \int_0^1 \int_0^1 x_v^2 x_u x_w g(x_v) g(x_u) g(x_w) dx_v dx_u dx_w}.$$

Substituting $\hat{x_i} = x_i^{1/a}$ for x_i where $i \in \{u, v, w\}$, we have that

$$= \frac{\int_0^1 \int_0^1 \int_0^1 \hat{x_v}^{2a} \hat{x_u}^{2a} \hat{x_w}^{2a} d\hat{x_v} d\hat{x_u} d\hat{x_w}}{\int_0^1 \int_0^1 \int_0^1 \hat{x_v}^{2a} \hat{x_u} \hat{x_w} d\hat{x_v} d\hat{x_u} d\hat{x_w}}$$

$$= \frac{(1+2a)^{-3}}{(1+2a)^{-1}(1+a)^{-2}} = \left(\frac{a+1}{2a+1}\right)^2.$$

QED.

So, for any vertices u, v, and w, the $P[u \sim v] < P[u \sim v | u \sim w \sim v]$, that is vertices are more likely to be adjacent if they share a common neighbor.

2.3.3 Diameter

In this section we show that if the vectors are iid some power of a uniform, then the Dense Random Dot Product Graph G almost surely consists of isolated vertices and a single giant connected component of diameter at most six. Before we prove this result, we establish the following lemmas.

Lemma 2.3.7. *Let $(G, \mathbf{X})$ be drawn from $\mathcal{D}[\mathbf{X}]$. Let $c = (\frac{\log n}{n})^{\frac{1}{4}}\omega_n$, where $\omega_n = \log\log n$. Let H be the subgraph of G induced by $V(H) = \{v \in V(G) : x_v \geq c\}$. Then $P[diam(H) > 2] \to 0$ as $n \to \infty$.*

Proof:

Take $u, v \in V(H)$ and $w \in V(H) - \{u, v\}$. Then $P[u \sim w \sim v] = x_u x_w^2 x_v$. Therefore, $P[\nexists \text{ a path } u \sim w \sim v] = 1 - x_u x_w^2 x_v \leq 1 - c^4$. Let $m = |V(H)|$ and $d(u, v)$ be the distance between u and v. Then

$$P[d(u,v) > 2] \leq P[\nexists\, w' \in V(H) - \{u, v\}, \text{ so that } u \sim w' \sim v]$$

$$= \prod_{w' \in V(H) - \{u,v\}} (1 - x_u x_{w'}^2 x_v)$$

$$\leq (1 - c^4)^{m-2}.$$

Now, for any vertex v, $P[x_v \geq c] = 1 - c^{\frac{1}{a}}$. Hence

$$E[|V(H)|] = n(1 - c^{\frac{1}{a}}) = n\left(1 - \left(\left(\frac{\log n}{n}\right)^{\frac{1}{4}} \omega_n\right)^{\frac{1}{a}}\right)$$

$$= n\left(1 - \left(\frac{\log n}{n}\right)^{\frac{1}{4a}} \omega_n^{\frac{1}{a}}\right).$$

So, for large n, $E[|V(H)|] = n - o(n)$. In addition, by Markov's inequality, $P[|V(H)| < 0.9n] = P[n - |V(H)| > 0.1n] \leq (nc^{1/a})/(0.1n) = 10c^{1/a} \to 0$. Therefore, with high probability

$$
\begin{aligned}
P[diam(H) > 2] &= \binom{m}{2} P[d(u,v) > 2] \\
&\leq \binom{m}{2} (1 - c^4)^{m-2} \\
&\leq \binom{n}{2} \left(1 - \left(\left(\frac{\log n}{n}\right)^{\frac{1}{4}} \omega_n\right)^4\right)^{0.9n-2} \\
&= \binom{n}{2} \left(1 - \left(\frac{\log n}{n}\right) (\omega_n)^4\right)^{0.9n-2} \\
&\sim n^2 e^{-0.9\omega_n^4 \log n} = \frac{n^2}{n^{0.9\omega_n^4}} \to 0
\end{aligned}
$$

as $n \to \infty$.

QED.

Given a graph G, a subgraph H of G, and a vertex $v \in V(G) - V(H)$, we write $v \not\sim H$ if for all $u \in V(H)$, $v \not\sim u$.

Lemma 2.3.8. *Let* $(G, \mathbf{X})$ *be drawn from* $\mathcal{D}[X]$ *and* H *be the subgraph of* G *as defined in Lemma 2.3.7. Let* $S = \{\{u, v\} : u, v \in V(G) - V(H), u \sim v,\ u \not\sim H, v \not\sim H\}$, *then* $E[|S|] \to 0$ *as* $n \to \infty$.

In other words, $P[|S| > 0] \to 0$ as $n \to \infty$.

Proof:

Let $u, v \in V(G) - V(H)$. Then $x_u, x_v < c$ and

$$P[u \sim v,\ u \not\sim H, v \not\sim H \mid x_u, x_v] = x_u x_v \prod_{y \in V(H)} (1 - x_u x_y)(1 - x_v x_y)$$

$$\leq x_u x_v \prod_{y \in V(H)\ :\ x_y \geq \frac{1}{2}} (1 - x_u x_y)(1 - x_v x_y)$$

$$\leq x_u x_v \prod_{y \in V(H)\ :\ x_y \geq \frac{1}{2}} \left(1 - \frac{1}{2} x_u\right)\left(1 - \frac{1}{2} x_v\right).$$

Now, for any vertex $w \in V(G)$, let $\beta := P[x_w \geq 1/2] = (1 - (\frac{1}{2})^{\frac{1}{a}})$. Therefore $E[|\{y\ :\ x_y \geq \frac{1}{2}\}|] = \beta n$. Also, by Chernoff's inequality, the number of such y is, with high probability, at least $0.9\beta n$

$$P[u \sim v,\ u \not\sim H, v \not\sim H | x_u, x_v] \leq x_u x_v \left(1 - \frac{1}{2} x_u\right)^{0.9\beta n} \left(1 - \frac{1}{2} x_v\right)^{0.9\beta n}.$$

Now we integrate to remove the conditioning on x_u, x_v and we have

$$P[u \sim v,\ u \not\sim H, v \not\sim H | x_u, x_v < c] =$$

$$= \frac{P[u \sim v,\ u \not\sim H, v \not\sim H, x_u, x_v < c]}{P[x_u, x_v < c]}$$

$$= \frac{1}{c^{\frac{2}{a}}} \int_0^c \int_0^c x_u x_v \left(1 - \frac{1}{2} x_u\right)^{0.9\beta n} \left(1 - \frac{1}{2} x_v\right)^{0.9\beta n} \frac{1}{a} x_u^{\frac{1-a}{a}}\, dx_u\ \frac{1}{a} x_v^{\frac{1-a}{a}}\, dx_v$$

$$= \frac{1}{c^{\frac{2}{a}}} \left(\int_0^c x \left(1 - \frac{1}{2}x\right)^{0.9\beta n} \frac{1}{a} x^{\frac{1-a}{a}} dx \right)^2.$$

Let $t = \sqrt[a]{nx}$, so that $dt = \frac{n^{\frac{1}{a}}}{a} x^{\frac{1-a}{a}} dx$ for $t \in (0, \sqrt[a]{nc})$ and substituting we have

$$P[u \sim v,\ u \not\sim H, v \not\sim H | x_u, x_v < c] = \frac{1}{c^{\frac{2}{a}}} \left(\int_0^{\sqrt[a]{nc}} \frac{t^a}{n} \left(1 - \frac{1}{2}\frac{t^a}{n}\right)^{0.9\beta n} \frac{dt}{n^{\frac{1}{a}}} \right)^2$$

$$= \frac{1}{c^{\frac{2}{a}} n^{2+\frac{2}{a}}} \left(\int_0^{\sqrt[a]{nc}} t^a \left(1 - \frac{1}{2}\frac{t^a}{n}\right)^{0.9\beta n} dt \right)^2$$

$$\sim \frac{1}{c^{\frac{2}{a}} n^{2+\frac{2}{a}}} \left(\int_0^{\infty} t^a \, e^{-\frac{0.9\beta t^a}{2}} dt \right)^2$$

$$= \frac{1}{c^{\frac{2}{a}} n^{2+\frac{2}{a}}} \left(\frac{2^{\frac{a+1}{a}} \Gamma(\frac{1}{a})}{(0.9\beta)^{\frac{a+1}{a}} a^2} \right)^2$$

$$= \Theta(1) \frac{1}{c^{\frac{2}{a}} n^{2+\frac{2}{a}}}.$$

Since $c = \left(\frac{\log n}{n}\right)^{\frac{1}{4}} \omega_n$ we have that

$$P[u \sim v,\ u \not\sim H, v \not\sim H | x_u, x_v < c] \le \Theta(1) \frac{1}{\left(\frac{\log n}{n}\right)^{\frac{1}{2a}} \omega_n^{\frac{2}{a}} n^{2+\frac{2}{a}}}$$

$$= \frac{\Theta(1)}{(\log n)^{\frac{1}{2a}} \omega_n^{\frac{2}{a}}} \frac{1}{n^{2+\frac{3}{2a}}}.$$

Now, for any vertex $v \in V(G)$, $P[x_v < c] = c^{\frac{1}{a}}$. Therefore

$$E[|V(G) - V(H)|] = nc^{\frac{1}{a}} = (\log n)^{\frac{1}{4a}} \omega_n^{\frac{1}{a}} \, n^{1-\frac{1}{4a}}.$$

So we have the following result

$$E[|S|] = \binom{E|V(G) - V(H)|}{2} P[u \sim v,\ u \not\sim H, v \not\sim H | x_u, x_v < c]$$

$$\leq (\log n)^{\frac{1}{2a}} \omega_n^{\frac{2}{a}}\ n^{2-\frac{1}{2a}} \frac{\Theta(1)}{(\log n)^{\frac{1}{2a}} \omega_n^{\frac{2}{a}}} \frac{1}{n^{2+\frac{3}{2a}}}$$

$$= \Theta(1) \frac{1}{n^{\frac{2}{a}}} \to 0$$

as $n \to \infty$.

QED.

Hence, with high probability, if $\exists\ u, v \in V(G) - V(H)$ with $u \sim v$ then either $u \sim H$ or $v \sim H$.

Now, as a result of Lemma 2.3.8, with high probability, any vertex that is not isolated is either adjacent to H or has a neighbor that is adjacent to H. Therefore with high probability, the probability that G is made up of one large component $\hat{G}$ and isolated vertices tends to one as n goes to ∞. With this in mind we obtain the following result.

Theorem 2.3.9. *Let* $(G, \mathbf{X})$ *be drawn from* $\mathcal{D}[X]$ *and let* $\hat{G}$ *be the subgraph of* G *induced by* $V(\hat{G}) = \{v \in V(G) : d(v) > 0\}$*. Then,* $P[diam(\hat{G}) \leq 6] \to 1$ *as* $n \to \infty$*.*

Proof:

Choose any two vertices $u, v \in V(\hat{G})$. Without loss of generality assume that $x_u \leq x_v$. We have the following three cases.

Case 1: $x_u, x_v \geq c$

By Lemma 2.3.7, we have $P[d(u,v) \leq 2] \to 1$ as $n \to \infty$.

Case 2: $x_u < c, x_v \geq c$

Now, since $\hat{G}$ is connected there exists $y \in \hat{G}$ such that $u \sim y$ and $d(u,v) \leq d(u,y) + d(y,v)$. If $x_y \geq c$ then by Lemma 2.3.7 as $n \to \infty$,

$$P[d(u,v) \leq 3] \geq P[d(y,v) \leq 3 - d(u,y)] = P[d(y,v) \leq 2] \to 1.$$

If $x_y < c$, then by Lemma 2.3.8 with probability $\to 1$ either u or y is adjacent to a vertex z with $x_z \geq c$. Hence

$$P[d(u,v) \leq 4] \geq P[d(u,z) + d(z,v) \leq 4]$$

$$= P[d(z,v) \leq 4 - d(u,z)]$$

$$\geq P[d(z,v) \leq 2] \to 1$$

as $n \to \infty$, by Lemma 2.3.7.

Case 3: $x_u, x_v < c$

Again, since $\hat{G}$ is connected there exists $y, z \in \hat{G}$ such that $u \sim y$ and $v \sim z$ and $d(u,v) \leq d(u,y) + d(y,z) + d(z,v) = 2 + d(y,z)$. Now, if $x_y, x_z \geq c$, then by Lemma

2.3.7, as $n \to \infty$,

$$P[d(u,v) \le 4] \ge P[d(y,z) \le 4 - d(u,y) - d(z,v)]$$

$$\ge P[d(y,z) \le 2] \to 1.$$

If $x_y < c$ and $x_z \ge c$. Then by Lemma 2.3.8 we know that there exists w with w adjacent to u or y and $x_w \ge c$. We have

$$P[d(u,v) \le 5] \ge P[d(u,w) + d(w,z) + d(z,v) \le 5]$$

$$= P[d(w,z) \le 5 - d(u,w) - d(w,z)]$$

$$\ge P[d(w,z) \le 2] \to 1$$

as $n \to \infty$ by Lemma 2.3.7. Likewise, if $x_y \ge c$ and $x_z < c$ then $P[d(u,v) \le 5] \to 1$ as $n \to \infty$. Finally if $x_y, x_z < c$, then by Lemma 2.3.8 there exists vertices w and t so that w is adjacent to u or y (i.e., $d(u,w) \le 2$) and t is adjacent to v or z (i.e., $d(v,t) \le 2$) with $x_w, x_t \ge c$. Therefore

$$P[d(u,v) \le 6] \ge P[d(u,w) + d(w,t) + d(t,v) \le 6]$$

$$= P[d(w,t) \le 6 - d(u,w) - d(t,v)]$$

$$\geq P[d(w,t) \leq 2] \to 1$$

as $n \to \infty$, by Lemma 2.3.7.

Therefore, for all cases as $n \to \infty$, $P[d(u,v) \leq 6] \to 1$ and so $P[diam(\hat{G}) \leq 6] \to 1$. QED.

2.4 Higher Dimensions

2.4.1 Basic Results in Higher Dimensions

The ideas presented above can be extended to higher dimensions. Suppose we draw each vector $\mathbf{x}_v$ from a distribution on $\mathbb{R}^t$, $t \in \mathbb{N}$. Additionally, we maintain the probability mapping f as the identity function. One possibility is to draw each $\mathbf{x}_v$ from $[\mathcal{U}^a[0, \frac{1}{2\sqrt[a]{t}}]]^t$ so that each component of $\mathbf{x}_v$ has density $g[(\mathbf{x}_v)_i] = \frac{t^{1/2a}}{a}(\mathbf{x}_v)_i^{\frac{a-1}{a}}$, and the density of $\mathbf{x}_v$ is $g(\mathbf{x}_v) = \prod g((\mathbf{x}_v)_i)$. Then, $\forall\, u, v \in V(G)$, $\mathbf{x}_u \cdot \mathbf{x}_v \in [0,1]$ and can be interpreted as a probability. We refer to this sample space of Dense Random Dot Product Graphs on n vertices in which t dimensional interest vectors are drawn from $[\mathcal{U}^a[0, \frac{1}{2\sqrt[a]{t}}]]^t$ as $\mathcal{D}[n,a,t]$. We have the following results.

Proposition 2.4.1. *Let G be drawn from $\mathcal{D}[n,a,t]$. For any $u, v \in V(G)$ the $P[u \sim v] = \frac{1}{(a+1)^2}$.*

Proof:

Let $\mathbf{x}$ and $\mathbf{y}$ be the vectors of vertices u and v, respectively. Then

$$P[u \sim v] = \int_{\left[0,\frac{1}{\sqrt{t}}\right]^t} \int_{\left[0,\frac{1}{\sqrt{t}}\right]^t} \mathbf{x} \cdot \mathbf{y}\, g(\mathbf{x}) g(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y}$$

$$= \int_0^{\frac{1}{\sqrt{t}}} \cdots \int_0^{\frac{1}{\sqrt{t}}} (x_1 y_1 + \cdots + x_t y_t) g(x_1) \cdots g(x_t) g(y_1) \cdots g(y_t) dx_1 \cdots dx_t dy_1 \cdots dy_t$$

(where x_i is the ith coordinate of $\mathbf{x}$, and y_i of $\mathbf{y}$)

$$= t^{\frac{t}{a}+1} \left[\frac{t^{\frac{-(a+1)}{2a}}}{a+1} t^{\frac{1-t}{2a}} \right]^2 = \frac{1}{(a+1)^2}$$

QED.

Hence, the expected number of edges is $\binom{n}{2}/(a+1)^2$ and is not dependent on the dimension from which we draw the interest vectors.

We can also ask questions about the degree distribution.

Theorem 2.4.2. *Let k be a fixed nonnegative integer. Let $\lambda(k)$ be the number of vertices of degree k in a random dot product graph drawn from $\mathcal{D}[n, a, t]$. Then as $n \to \infty$*

$$E[\lambda(k)] \sim C(k, t, a) n^{\frac{a-t}{a}}$$

where $C(k, t, a)$ is a constant depending only on $k, t,$ and a.

Proof:

Choose $v \in V(G)$ fixed, but arbitrary, and the denote the vector of v by $\mathbf{y}$. Let $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n-1)}$ be the vectors of the remaining $n-1$ vertices in $V(G) - \{v\}$. Then, without loss of generality, the conditional probability

$$P[d(v) = k|\mathbf{y}] = \binom{n}{k} \int_{\left[0,\frac{1}{\sqrt{t}}\right]^t} \cdots \int_{\left[0,\frac{1}{\sqrt{t}}\right]^t} (\mathbf{x}^{(1)}\cdot\mathbf{y}) \cdots (\mathbf{x}^{(k)}\cdot\mathbf{y})(1-\mathbf{x}^{(k+1)}\cdot\mathbf{y}) \cdots (1-\mathbf{x}^{(n)}\cdot\mathbf{y})$$

$$\cdot\, g(\mathbf{x}^{(1)}) \cdots g(\mathbf{x}^{(n-1)})\, d\mathbf{x}^{(1)} \cdots d\mathbf{x}^{(n-1)}$$

which is separable. For each of the $\mathbf{x}^{(i)} \cdot \mathbf{y}$ terms

$$\int_{\left[0,\frac{1}{\sqrt{t}}\right]^t} (\mathbf{x}^{(i)} \cdot \mathbf{y})\, g(\mathbf{x}^{(i)})\, d\mathbf{x}^{(i)} =$$

$$= \int_0^{\frac{1}{\sqrt{t}}} \cdots \int_0^{\frac{1}{\sqrt{t}}} (x_1^{(i)} y_1 + \cdots + x_t^{(i)} y_t) \left(\frac{\sqrt[2a]{t}}{a} (x_1^{(i)})^{\frac{1}{a}-1} \right) \cdots \left(\frac{\sqrt[2a]{t}}{a} (x_t^{(i)})^{\frac{1}{a}-1} \right) dx_1^{(i)} \cdots dx_t^{(i)}$$

$$= \frac{y_1 + \cdots + y_t}{\sqrt{t}(a+1)}.$$

Likewise for each of the $1 - \mathbf{x}^{(i)} \cdot \mathbf{y}$ terms

$$\int_{\left[0,\frac{1}{\sqrt{t}}\right]^t} (1 - \mathbf{x}^{(i)} \cdot \mathbf{y})\, g(\mathbf{x}^{(i)})\, d\mathbf{x}^{(i)} =$$

$$= \int_0^{\frac{1}{\sqrt{t}}} \cdots \int_0^{\frac{1}{\sqrt{t}}} (1-(x_1^{(i)} y_1 + \cdots + x_t^{(i)} y_t)) \left(\frac{\sqrt[2a]{t}}{a} (x_1^{(i)})^{\frac{1}{a}-1} \right) \cdots \left(\frac{\sqrt[2a]{t}}{a} (x_t^{(i)})^{\frac{1}{a}-1} \right) dx_1^{(i)} \cdots dx_t^{(i)}$$

$$= 1 - \frac{y_1 + \cdots + y_t}{\sqrt{t}(a+1)}.$$

Therefore, we have that

$$P[d(v) = k|\mathbf{y}] = \binom{n}{k} \left(\frac{y_1 + \cdots + y_t}{\sqrt{t}(a+1)}\right)^k \left(1 - \frac{y_1 + \cdots + y_t}{\sqrt{t}(a+1)}\right)^{n-k}.$$

Now, we remove the conditioning on $\mathbf{y}$

$$P[d(v) = k] =$$

$$= \binom{n}{k} \int_0^{\frac{1}{\sqrt{t}}} \cdots \int_0^{\frac{1}{\sqrt{t}}} \left(\frac{y_1 + \cdots + y_t}{\sqrt{t}(a+1)}\right)^k \left(1 - \frac{y_1 + \cdots + y_t}{\sqrt{t}(a+1)}\right)^{n-k} \left(\frac{\sqrt[2a]{t}}{a} y^{\frac{1}{a}-1}\right)$$

$$\cdots \left(\frac{\sqrt[2a]{t}}{a} y_t^{\frac{1}{a}-1}\right) dy_1 \cdots dy_t.$$

Letting each $\hat{y}_i = \sqrt[a]{n}\, y_i^{1/a}$,we have that $d\hat{y}_i = \sqrt[a]{n}\, y_i^{\frac{1}{a}-1}\, dy_i$ and $0 \leq \hat{y}_i \leq \frac{\sqrt[a]{n}}{\sqrt[2a]{t}}$.

Therefore, substituting we have

$$= \binom{n}{k} \int_0^{\frac{\sqrt[a]{n}}{\sqrt[2a]{t}}} \cdots \int_0^{\frac{\sqrt[a]{n}}{\sqrt[2a]{t}}} \left(\frac{\hat{y}_1^a + \cdots + \hat{y}_t^a}{\sqrt{t}n(a+1)}\right)^k \left(1 - \frac{\hat{y}_1^a + \cdots + \hat{y}_t^a}{\sqrt{t}n(a+1)}\right)^{n-k} \left(\frac{\sqrt[2a]{t}}{\sqrt[a]{n}}\right)^t d\hat{y}_1 \cdots d\hat{y}_t$$

which as $n \to \infty$,

$$\sim \binom{n}{k} \left(\frac{\sqrt[2a]{t}}{\sqrt[a]{n}}\right)^t \left(\frac{1}{n^d}\right) \int_0^{\infty} \cdots \int_0^{\infty} \left(\frac{\hat{y}_1^a + \cdots + \hat{y}_t^a}{\sqrt{t}(a+1)}\right)^k \exp\left\{-\frac{\hat{y}_1^a + \cdots + \hat{y}_t^a}{\sqrt{t}(a+1)}\right\} d\hat{y}_1 \cdots d\hat{y}_t.$$

Noting that the the integration is a constant with respect to n, we have that

$$P[d(v) = k] \sim \binom{n}{k} \frac{1}{n^{k+\frac{t}{a}}} C(k,t,a) \sim \frac{1}{n^{\frac{t}{a}}} C(k,t,a).$$

Therefore as $n \to \infty$ the expected number of vertices of degree k is

$$E[\lambda(k)] = n\, P[d(v) = k] \sim C(k,t,a)\, n^{1-\frac{t}{a}}.$$

QED.

2.4.2 A Bend in the Power Law

We would like to demonstrate that realizations of $\mathcal{D}[n,a,t]$ obey the degree distribution power law, for all $a > t$. However, in dimensions $t > 1$ empirical data appears to contain a bend in the log-log plot of the degree distribution. Therefore, we believe that $\mathcal{D}[n,a,t]$ does not strictly obey the power law when $t > 1$. Interestingly, we have observed various real world data sets in which this is also the case [19].

To understand why this bend occurs, we examine the simpler case of $\mathcal{D}[n,a,2]$. Let us consider the conditional expected degree of a vertex where we condition on the vector assigned to the vertex.

Let $z \in V(G)$ have vector $\mathbf{z} = [z_1, z_2]^T$. Then

$$E[d(z)|\mathbf{z}] = E\left[\sum_{w \in V(G)\,:\,w \neq z} \mathbf{I}\{v \sim w|\mathbf{z}\}\right]$$

$$= (n-1)\,P[z \sim w|\mathbf{z}]$$

$$= (n-1)\int_0^{\frac{1}{\sqrt{2}}}\int_0^{\frac{1}{\sqrt{2}}}(z_1x_1 + z_2x_2)g(x_1)g(x_2)dx_1dx_2$$

$$= \frac{n-1}{a+1}(z_1+z_2)\frac{1}{\sqrt{2}} \sim \frac{n}{(a+1)\sqrt{2}}||\mathbf{z}||_1$$

Because $d(z)$ is the sum of $n-1$ iid Bernouli random variables, its distribution is highly concentrated about its mean (assuming $d(z)$ is large). So, more or less, the degree of a vertex follows the sum of the coordinates of its representing vector.

Thus, the ℓ_1-norm of $\mathbf{z}$ is a sentinel for $d(z)$. The number of vertices of degree at most k should be around

$$nP\left[||\mathbf{x}||_1 \leq \frac{k\sqrt{2}(a+1)}{n}\right].$$

With this in mind, we define $f(b) = P\Big[||\mathbf{x}||_1 \leq b\Big]$ where $0 \leq b \leq \frac{2}{\sqrt{2}}$. We calculate this probability as follows.

First, note that since $\mathbf{x} \sim [\mathcal{U}^a[0, \frac{1}{\sqrt[2a]{2}}]]^2$ we have that

$$P[x_i \leq s] = P\left[Y \leq s^{1/a}\right] = s^{1/a}\,\sqrt[2a]{2}.$$

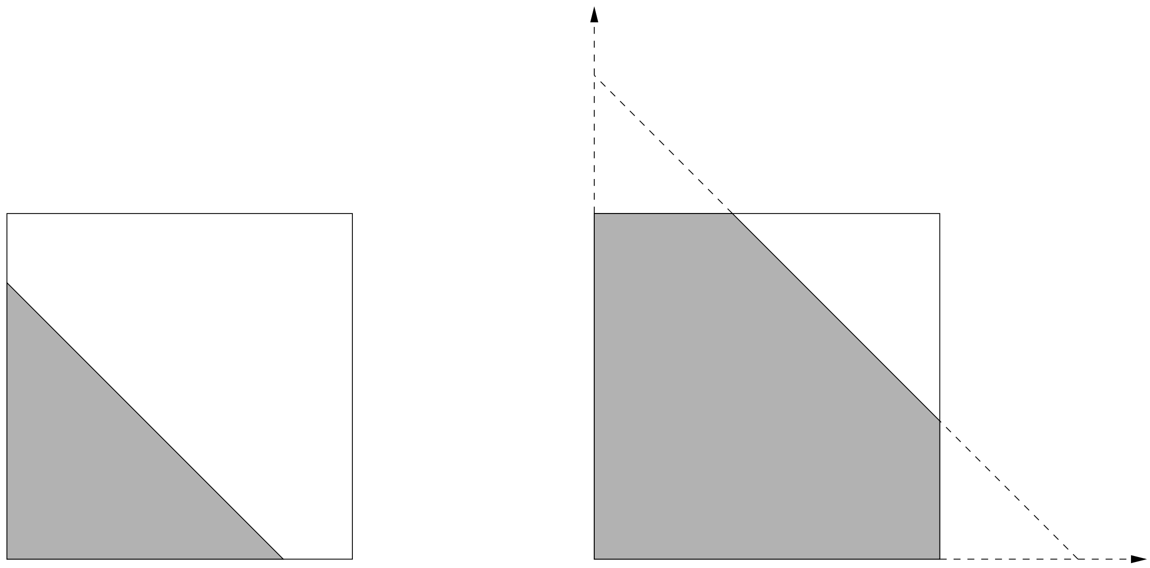


Figure 2.1: The domain $\mathcal{U}(b)$

where $Y \sim \mathcal{U}[0, \frac{1}{\sqrt[2a]{2}}]$. Therefore, the density function is

$$g(s) = \frac{d}{ds}(s^{1/a}\sqrt[2a]{2}) = \frac{s^{-1+\frac{1}{a}}\sqrt[2a]{2}}{a}.$$

Therefore,

$$f(b) = P[||\mathbf{x}||_1 \leq b] = \int\int_{\mathcal{U}(b)} g(x)g(y)dxdy$$

where $\mathcal{U}(b)$ is the domain $\{(x,y) : 0 \leq x, y \leq \frac{1}{\sqrt{2}}, x + y \leq b\}$. This domain is illustrated by the diagrams in Figure 2.1. The diagram on the left shows $\mathcal{U}(b)$ in the case $0 \leq b \leq \frac{1}{\sqrt{2}}$ and the diagram on the right for $\frac{1}{\sqrt{2}} \leq b \leq \frac{2}{\sqrt{2}}$.

In case $0 \leq b \leq \frac{1}{\sqrt{2}}$,

$$f(b) = \int_0^b \int_0^{b-x} g(x)g(y)dxdy = \frac{\sqrt{\pi}\sqrt[a]{2}\Gamma(1+\frac{1}{a})}{4^{1/a}\Gamma(\frac{1}{2}+\frac{1}{a})}b^{2/a} = K_a b^{2/a}$$

where K_a is a constant that depends only on a.

The situation for $\frac{1}{\sqrt{2}} \leq b \leq \frac{2}{\sqrt{2}}$ is more complex. It is simpler to integrate over the complementary domain $[0, \frac{1}{\sqrt{2}}] - \mathcal{U}(b)$ (the right triangle in the upper right corner in the right hand diagram in Figure 2.1) and subtract the result from 1. Doing this we obtain the following:

$$f(b) = 1 - \int_{b-\frac{1}{\sqrt{2}}}^{\frac{1}{\sqrt{2}}} \int_{b-x}^{\frac{1}{\sqrt{2}}} g(y)g(x)dydx$$

$$= \frac{(b-\frac{1}{\sqrt{2}})^{1/a}}{\frac{1}{\sqrt[2a]{2}}} - \frac{b^{1/a}(b-\frac{1}{\sqrt{2}})^{1/a}}{\frac{1}{\sqrt[a]{2}}} F\left(\begin{array}{c|c} \frac{1}{a}, -\frac{1}{a} & \frac{b-\frac{1}{\sqrt{2}}}{b} \\ \frac{a+1}{a} & \end{array}\right) + \frac{b^{1/a}}{\frac{1}{\sqrt[2a]{2}}} F\left(\begin{array}{c|c} \frac{1}{a}, -\frac{1}{a} & \frac{\frac{1}{\sqrt{2}}}{b} \\ \frac{a+1}{a} & \end{array}\right)$$

where F is the $2,1$-hypergeometric function.

Some further comments can be made. For $0 \leq b \leq \frac{1}{\sqrt{2}}$, the distribution behaves as a power of b, and this gives rise to a linear regime in the $\log\log$ histogram of the degree distribution. The number of vertices of degree at most k is a constant times $k^{2/a}$, so the histogram will have slope $\frac{2}{a} - 1$ (derivative with respect to b, or equivalently, k).

Then, at $b = \frac{1}{\sqrt{2}}$, which corresponds to $k = \frac{n}{2(t+1)}$, there will be a bend in the distribution, and for $b > \frac{1}{\sqrt{2}}$ the behavior of the distribution (in the $\log\log$ plot) will not be linear.

We believe that this can be extended to higher dimensions as follows:

Conjecture 2.4.3. *Let G be drawn from $\mathcal{D}[n, a, t]$, $t > 1$. Then for all $a > t$, the* log-log *plot of the degree distribution of* G

- *obeys the power law for degrees* $k \in [1, \frac{n}{(a+1)t}]$ *and will have slope* $\frac{t-a}{a}$ *and*

- *has an initial bend in the plot occurring at the degree* $k = \frac{n}{(a+1)t}$.

Chapter 3

The Sparse Model

In this chapter we present the Sparse Random Dot Product Graph. We begin by introducing the model and proving basic results in Section 3.1.1. The main results parallel those in the dense case and are presented in Section 3.2. In Section 3.2.2 we show that the degree distribution obeys a power law. In Section 3.2.3 we show that the Sparse Random Dot Product graph has a small diameter. Additionally, unlike the dense model in which all small subgraphs appear with probability tending to one, in the sparse model subgraphs appear at certain thresholds dependent upon a parameter b that is not used in the dense model. In Sections 3.1.1 and 3.1.3 we present specific results regarding the thresholds for the appearance of edges, cliques, cycles, and trees. Then in Section 3.2.1, we prove a general threshold result for the appearance of any fixed graph H. Finally, in Section 3.3 we recap our results, discuss the evolution of the Sparse Random Dot Product Graph as b goes from zero to infinity.

3.1 The Sparse Random Dot Product Graph

In this chapter, we investigate a version of the Random Dot Product Graph in which the probability mapping f is not the identity function. Instead let $f(r) = \frac{r}{n^b}$ where b is a positive[1] real number. We denote this sample space of Sparse Random Dot Product Graphs on n vertices in which one dimensional vectors are drawn from $\mathcal{U}^a[0,1]$ $(a > 1)$ and for which the probability mapping is $f(r) = \frac{r}{n^b}$ as $\mathcal{DS}[n, a, b, 1]$.

3.1.1 Results in the General Case: $b \in (0, \infty)$

We begin by studying the model for the general case when $b \in (0, \infty)$. We have the following basic results.

Proposition 3.1.1. *Let G be drawn from $\mathcal{DS}[n, a, b, 1]$ with $b \in (0, \infty)$. For any $u, v \in V(G)$ we have*

$$P[u \sim v] = \frac{1}{n^b\,(a+1)^2}.$$

Proof:

$$P[u \sim v] = \int_0^1 \int_0^1 \frac{x_u\,x_v}{n^b}\,g(x_u)\,g(x_v)\,dx_u\,dx_v = \frac{1}{n^b\,a^2} \int_0^1 \int_0^1 x_u^{1/a}\,x_v^{1/a}\,dx_u\,dx_v$$

$$= \frac{1}{n^b\,(a+1)^2}.$$

[1]Note that the case $b = 0$ is the model discussed in Chapter 2.

QED.

Thus, an arbitrary edge appears in the graph with probability $\frac{1}{n^b\,(a+1)^2}$ and the expected number of edges is $\binom{n}{2}\frac{1}{n^b\,(a+1)^2} \asymp \frac{n^{2-b}}{(a+1)^2}$. We would like to know for which values of b does a Sparse Random Dot Product Graph have edges (with high probability) and so first we calculate the variance on the number of edges.

Lemma 3.1.2. *Let G be drawn from $\mathcal{DS}[n, a, b, 1]$ with $b \in (0, \infty)$. Let the random variable $\mathbf{X}$ be the number of edges in G. Then*

$$Var[\mathbf{X}] = \frac{\binom{n}{2}}{n^b(a+1)^2} + \frac{6\binom{n}{3}}{n^{2b}(1+2a)(a+1)^2} + \frac{\binom{n}{2}\binom{n-2}{2}}{n^{2b}(a+1)^4} - \frac{\binom{n}{2}^2}{n^{2b}(a+1)^4}.$$

Proof:

Consider $\mathbf{X}$ as the sum of the indicator functions for each individual edge. So, $\mathbf{X} = \sum_{\substack{u,v \in V(G) \\ u<v}} \mathbf{I}\{u \sim v\}$. Therefore by Proposition 3.1.1 we have that

$$E[\mathbf{X}] = E\left[\sum_{\substack{u,v \in V(G) \\ u<v}} \mathbf{I}\{u \sim v\}\right] = \sum_{\substack{u,v \in V(G) \\ u<v}} E[\mathbf{I}\{u \sim v\}] = \sum_{\substack{u,v \in V(G) \\ u<v}} P[u \sim v] = \frac{\binom{n}{2}}{n^b(a+1)^2}.$$

The variance of $\mathbf{X}$ is

$$\mathrm{Var}[\mathbf{X}] = E[\mathbf{X}^2] - E[\mathbf{X}]^2 = E[\mathbf{X}^2] - \left(\frac{\binom{n}{2}}{n^b(a+1)^2}\right)^2$$

and so we calculate $E[\mathbf{X}^2]$.

$$E[\mathbf{X}^2] = E\left[\left(\sum_{\substack{u,v\in V(G)\\u<v}} \mathbf{I}\left\{u \sim v\right\}\right)^2\right]$$

$$= E\left[\sum_{\substack{u,v\in V(G)\\u<v}} \mathbf{I}^2\left\{u \sim v\right\} + \sum_{\substack{u,v,w\in V(G)\\u\neq v,u\neq w,v\neq w}} \mathbf{I}\left\{u \sim v\right\}\mathbf{I}\left\{v \sim w\right\}\right.$$

$$\left. + \sum_{\substack{u,v,w,z\in V(G)\\u<v,w<z,\{u,v\}\cap\{w,z\}=\emptyset}} \mathbf{I}\left\{u \sim v\right\}\mathbf{I}\left\{w \sim z\right\}\right]$$

$$= E\left[\sum_{\substack{u,v\in V(G)\\u<v}} \mathbf{I}^2\left\{u \sim v\right\}\right] + E\left[\sum_{\substack{u,v,w\in V(G)\\u\neq v,u\neq w,v\neq w}} \mathbf{I}\left\{u \sim v\right\}\mathbf{I}\left\{v \sim w\right\}\right]$$

$$+ E\left[\sum_{\substack{u,v,w,z\in V(G)\\u<v,w<z,\{u,v\}\cap\{w,z\}=\emptyset}} \mathbf{I}\left\{u \sim v\right\}\mathbf{I}\left\{w \sim z\right\}\right].$$

We now look at each of the three expected values individually. In the first expected value, since the indicator random variables can only have the value of 0 or 1 we see that

$$E\left[\sum_{\substack{u,v\in V(G)\\u<v}} \mathbf{I}^2\left\{u \sim v\right\}\right] = E\left[\sum_{\substack{u,v\in V(G)\\u<v}} \mathbf{I}\left\{u \sim v\right\}\right] = \sum_{\substack{u,v\in V(G)\\u<v}} E\left[\mathbf{I}\left\{u \sim v\right\}\right]$$

$$= \binom{n}{2} P[u \sim v] = \frac{\binom{n}{2}}{n^b(a+1)^2} = E[\mathbf{X}].$$

Next, for the middle term, we have that

$$E\left[\sum_{\substack{u,v,w\in V(G)\\u\neq v,u\neq w,v\neq w}} \mathbf{I}\{u\sim v\}\,\mathbf{I}\{v\sim w\}\right] = \sum_{\substack{u,v,w\in V(G)\\u\neq v,u\neq w,v\neq w}} E\left[\mathbf{I}\{u\sim v\}\,\mathbf{I}\{v\sim w\}\right]$$

$$= 6\binom{n}{3}E\left[\mathbf{I}\{u\sim v\}\,\mathbf{I}\{v\sim w\}\right].$$

Now, $\mathbf{I}\{u\sim v\}\,\mathbf{I}\{v\sim w\} = 1$ if and only if $\mathbf{I}\{u\sim v\} = 1$ and $\mathbf{I}\{v\sim w\} = 1$, and is 0 otherwise. Therefore,

$$E\left[\mathbf{I}\{u\sim v\}\,\mathbf{I}\{v\sim w\}\right] = P[u\sim v\sim w] = \int_0^1\int_0^1\int_0^1 \frac{x_u^a\,x_v^a}{n^b}\frac{x_v^a\,x_w^a}{n^b}dx_u\,dx_v\,dx_w$$

$$= \frac{1}{n^{2b}(1+2a)(1+a)^2}.$$

Hence the middle term is

$$E\left[\sum_{\substack{u,v,w\in V(G)\\u\neq v,u\neq w,v\neq w}} \mathbf{I}\{u\sim v\}\,\mathbf{I}\{v\sim w\}\right] = \frac{6\binom{n}{3}}{n^{2b}(1+2a)(1+a)^2}.$$

Finally, since the events $\{u\sim v\}$ and $\{w\sim z\}$ are independent, the last term becomes

$$E\left[\sum_{\substack{u,v,w,z\in V(G)\\u<v,w<z\\\{u,v\}\cap\{w,z\}=\emptyset}} \mathbf{I}\{u\sim v\}\,\mathbf{I}\{w\sim z\}\right] = \sum_{\substack{u,v,w,z\in V(G)\\u<v,w<z\\\{u,v\}\cap\{w,z\}=\emptyset}} E\left[\mathbf{I}\{u\sim v\}\right]E\left[\mathbf{I}\{w\sim z\}\right]$$

$$= \binom{n}{2}\binom{n-2}{2} P[u \sim v]P[w \sim z] = \binom{n}{2}\binom{n-2}{2}\left(\frac{1}{n^b(a+1)^2}\right)^2.$$

Now we have that $E[\mathbf{X}^2] = \frac{\binom{n}{2}}{n^b(a+1)^2} + \frac{6\binom{n}{3}}{n^{2b}(1+2a)(a+1)^2} + \frac{\binom{n}{2}\binom{n-2}{2}}{n^{2b}(a+1)^4}$. Therefore the variance is indeed

$$\mathrm{Var}[\mathbf{X}] = \frac{\binom{n}{2}}{n^b(a+1)^2} + \frac{6\binom{n}{3}}{n^{2b}(1+2a)(a+1)^2} + \frac{\binom{n}{2}\binom{n-2}{2}}{n^{2b}(a+1)^4} - \frac{\binom{n}{2}^2}{n^{2b}(a+1)^4}.$$

QED.

Theorem 3.1.3. *Let G be drawn from $\mathcal{DS}[n, a, b, 1]$ with $b \in (0, \infty)$. A threshold for the appearance of edges in G is $b = 2$.*

Proof:

Let the random variable $\mathbf{X}$ be the number of edges in G. Then we know from the proof of Lemma 3.1.2 that $E(\mathbf{X}) = \frac{\binom{n}{2}}{n^b(a+1)^2}$. Let $b = 2$ be our candidate for the threshold function for the appearance of edges. Now, by Markov's inequality $P[\mathbf{X} \geq 1] \leq E[\mathbf{X}] = \frac{\binom{n}{2}}{n^b(a+1)^2} \sim \frac{n^{2-b}}{(a+1)^2}$. So whenever $b > 2$, we have that $E[\mathbf{X}] \to 0$ and therefore the $P[\mathbf{X} = 0] \to 1$ as $n \to \infty$ and almost every graph has no edges.

Also, whenever, $b < 2$, then $E[\mathbf{X}] \to \infty$ and we use the second moment method. By Chebychev's inequality we know that $P[\mathbf{X} = 0] \leq \mathrm{Var}(\mathbf{X})/E[\mathbf{X}]^2$. By Lemma 3.1.2 we have that

$$P[\mathbf{X} = 0] \leq \frac{\mathrm{Var}[\mathbf{X}]}{E[\mathbf{X}]^2} = \frac{\frac{\binom{n}{2}}{n^b(a+1)^2} + \frac{6\binom{n}{3}}{n^{2b}(1+2a)(a+1)^2} + \frac{\binom{n}{2}\binom{n-2}{2}}{n^{2b}(a+1)^4} - \frac{\binom{n}{2}^2}{n^{2b}(a+1)^4}}{\frac{\binom{n}{2}^2}{n^{2b}(a+1)^4}}.$$

Now let us look at each piece of the fraction separately. The first term is

$$\frac{\frac{\binom{n}{2}}{n^b(a+1)^2}}{\frac{\binom{n}{2}^2}{n^{2b}(a+1)^4}} \sim \frac{n^2 n^{2b}(a+1)^4}{n^2 n^b(a+1)^2} = \frac{n^b(a+1)^2}{n^2} \to 0$$

since $b < 2$. Similarly the second term is

$$\frac{\frac{6\binom{n}{3}}{n^{2b}(1+2a)(a+1)^2}}{\frac{\binom{n}{2}^2}{n^{2b}(a+1)^4}} \sim \frac{6\,n^3 n^{2b}(a+1)^4}{n^4 n^{2b}(1+2a)(1+a)^2} \to 0.$$

Finally the last term is

$$\frac{\frac{\binom{n}{2}\binom{n-2}{2}}{n^{2b}(a+1)^4} - \frac{\binom{n}{2}^2}{n^{2b}(a+1)^4}}{\frac{\binom{n}{2}^2}{n^{2b}(a+1)^4}} < 0.$$

Therefore $P[\mathbf{X} = 0] \to 0$ whenever $b < 2$ and so $b = 2$ is a threshold for the appearance of edges.

QED.

So, the threshold for the appearance of edges is $b = 2$ or when the probability mapping is $f(r) = \frac{r}{n^2}$. Therefore whenever $b \in (2, \infty)$ a Sparse Random Dot Product Graph is almost surely edgeless. It is interesting to note that $1/n^2$ is a threshold function for the appearance of edges in a Erdős-Rényi random graph. Next we examine the threshold for the appearance of a clique of size $k \geq 2$.

Proposition 3.1.4. *Let G be drawn from $\mathcal{DS}[n, a, b, 1]$ with $b \in (0, \infty)$. Let $k \in \mathbb{N}, k > 1$. For any distinct $u_1, u_2, \ldots, u_k \in V(G)$ we have that the probability G has a* *clique of size k on $u_1, u_2, \ldots, u_k$ is*

$$P[K_k \text{ on } u_1, u_2, \ldots, u_k] = \frac{(1+(k-1)a)^{-k}}{n^{\binom{k}{2}b}}.$$

Proof:

$$P[K_k \text{ on } u_1, u_2, \ldots, u_k] = \int_0^1 \cdots \int_0^1 \frac{x_{u_1}\, x_{u_2}}{n^b} \frac{x_{u_1}\, x_{u_3}}{n^b} \cdots \frac{x_{u_1}\, x_{u_k}}{n^b} \cdots \frac{x_{u_{k-1}}\, x_{u_k}}{n^b}$$

$$g(x_{u_1})\, g(x_{u_2}) \,\cdots g(x_{u_k})\, dx_{u_1}\, dx_{u_2} \cdots dx_{u_k}$$

$$= \frac{1}{n^{\binom{k}{2}b}} \int_0^1 \cdots \int_0^1 x_{u_1}^{k-1}\, x_{u_2}^{k-1} \cdots x_{u_k}^{k-1} g(x_{u_1})\, g(x_{u_2}) \,\cdots g(x_{u_k})\, dx_{u_1}\, dx_{u_2} \cdots dx_{u_k}$$

$$= \frac{1}{n^{\binom{k}{2}b}} \left[\int_0^1 x_{u_i}^{k-1} g(x_{u_i}) dx_{u_i} \right]^k$$

$$= \frac{(1+(k-1)a)^{-k}}{n^{\binom{k}{2}b}}.$$

QED.

Therefore the expected number of cliques of size k is $\binom{n}{k} \frac{(1+(k-1)a)^{-k}}{n^{\binom{k}{2}b}}$.

Theorem 3.1.5. *Let G be drawn from $\mathcal{DS}[n, a, b, 1]$ with $b \in (0, \infty)$. Let $k \in \mathbb{N}, k > 1$. A threshold for the appearance of cliques of size k is $b = \frac{2}{k-1}$.*

Proof:

Let the random variable $\mathbf{X}$ be the number of cliques of size k in G. Then we know by Proposition 3.1.4 that $E(\mathbf{X}) = \binom{n}{k} \frac{(1+(k-1)a)^{-k}}{n^{\binom{k}{2}b}}$. Let $b = \frac{2}{k-1}$ be our candidate for the threshold for the appearance of K_k. Now, by Markov's inequality

$$P[\mathbf{X} \geq 1] \leq E[\mathbf{X}] = \binom{n}{k} \frac{(1+(k-1)a)^{-k}}{n^{\binom{k}{2}b}} \sim \frac{n^{k-\frac{k(k-1)}{2}b}}{(1+(k-1)a)^k}.$$

So whenever $b > \frac{2}{k-1}$, we have that $E[\mathbf{X}] \to 0$ and therefore the $P[\mathbf{X} = 0] \to 1$ as $n \to \infty$ and almost every graph has no clique of size k.

Also, whenever, $b < \frac{2}{k-1}$, then $E[\mathbf{X}] \to \infty$ and we apply the second moment method. By Chebychev's inequality we know that $P[\mathbf{X} = 0] \leq \mathrm{Var}(\mathbf{X})/E[\mathbf{X}]^2$.

Let $I_{u_1,u_2,\ldots,u_k}$ be the indicator function for when the vertices $u_1, u_2, \ldots, u_k$ are a clique. Then $\mathbf{X} = \sum_{u_1<u_2<\cdots<u_k} I_{u_1,u_2,\ldots,u_k}$. The variance of $\mathbf{X}$ is

$$\mathrm{Var}[\mathbf{X}] = E[\mathbf{X}^2] - E[\mathbf{X}]^2 = E[\mathbf{X}^2] - \left[\binom{n}{k} \frac{(1+(k-1)a)^{-k}}{n^{\binom{k}{2}b}}\right]^2$$

and so we calculate $E[\mathbf{X}^2]$.

$$E[\mathbf{X}^2] = E\left[\left(\sum_{u_1<u_2<\cdots<u_k} I_{u_1,u_2,\ldots,u_k}\right)^2\right]$$

$$= E\left[\left(\sum_{u_1<u_2<\cdots<u_k} I_{u_1,u_2,\ldots,u_k}\right)\left(\sum_{y_1<y_2<\cdots<y_k} I_{y_1,y_2,\ldots,y_k}\right)\right]$$

$$= \sum_{\substack{u_1<u_2<\cdots<u_k \\ y_1<y_2<\cdots<y_k}} E[I_{u_1,u_2,\ldots,u_k} I_{y_1,y_2,\ldots,y_k}].$$

Now in each term of the summation $u_1, \ldots, u_k$ and $y_1, \ldots y_k$ overlap in up to k terms. Let $\{z_1, z_2, \ldots, z_{k+j}\} = \{u_1, \ldots, u_k, y_1, \ldots y_k\}$ then $u_1, \ldots, u_k$ and $y_1, \ldots y_k$

overlap in exactly $2k-(k+j)=k-j$ terms, for some j with $0 \leq j \leq k$, and we can rewrite the summation as

$$E[\mathbf{X}^2] = \sum_{j=0}^{k} \sum_{\substack{u_1<u_2<\cdots<u_k \\ y_1<y_2<\cdots<y_k \\ |\{u_1,\ldots,u_k,y_1,\ldots y_k\}|=k+j}} E[I_{u_1,u_2,\ldots,u_k} I_{y_1,y_2,\ldots,y_k}]$$

$$= \sum_{j=0}^{k} \sum_{\substack{u_1<u_2<\cdots<u_k \\ y_1<y_2<\cdots<y_k \\ |\{u_1,\ldots,u_k,y_1,\ldots y_k\}|=k+j}} P[u_1,u_2,\ldots,u_k \text{ is a clique and } y_1,\ldots,y_k \text{ is a clique}].$$

Let us consider this probability for a moment. In this integral, each edge appears only once, regardless if it occurs in both cliques. Therefore there are exactly $2\binom{k}{2} - \binom{k-j}{2}$ edges each contributing a $\frac{1}{n^b}$ term to the integral. Also, without loss of generality, let us assume that $z_1,\ldots,z_{k-j}$ are the vertices that overlap, i.e. they appear in both cliques. Therefore, they each appear in exactly $k-1+j$ edges, while the remaining vertices appear in only $k-1$ edges. Hence, their vectors, $x_{z_1},\ldots,x_{z_{k-j}}$ appear in the integral $k-1+j$ times and the vectors $x_{z_{k-j+1}},\ldots x_{z_{k+j}}$ appear $k-1$ times. So we see that the probability is

$$P[u_1,u_2,\ldots,u_k \text{ is a clique and } y_1,\ldots,y_k \text{ is a clique}] =$$

$$= \frac{1}{n^{[2\binom{k}{2}-\binom{k-j}{2}]b}} \int_0^1 \cdots \int_0^1 x_{z_1}^{a(k+j-1)} \cdots x_{z_{k-j}}^{a(k+j-1)} x_{z_{k-j+1}}^{a(k-1)} \cdots x_{z_{k+j}}^{a(k-1)} dx_{z_1} \cdots dx_{z_{k+j}}$$

which by separating the integrals and collecting like terms becomes

$$= \frac{1}{n^{\left[2\binom{k}{2}-\binom{k-j}{2}\right]b}} \left[\int_0^1 x_{z_{k-j+1}}^{a(k-1)} dx_{z_{k-j+1}}\right]^{2j} \left[\int_0^1 x_{z_1}^{a(k-1+j)} dx_{z_1}\right]^{k-j}$$

$$= \frac{1}{n^{\left[2\binom{k}{2}-\binom{k-j}{2}\right]b}} \left[\frac{1}{a(k-1)+1}\right]^{2k} \left[\frac{1}{a(k-1+j)+1}\right]^{k-j}.$$

So, we see that our equation becomes

$$E[\mathbf{X}^2] = \sum_{j=0}^{k} \sum_{\substack{u_1<u_2<\cdots<u_k \\ y_1<y_2<\cdots<y_k \\ |\{u_1,\ldots,u_k,y_1,\ldots y_k\}|=k+j}} \frac{1}{n^{\left[2\binom{k}{2}-\binom{k-j}{2}\right]b}} \left[\frac{1}{a(k-1)+1}\right]^{2j} \left[\frac{1}{a(k-1+j)+1}\right]^{k-j}$$

$$= \sum_{j=0}^{k} \frac{\binom{k}{k-j}\binom{n}{k}\binom{n-k}{j}}{n^{\left[2\binom{k}{2}-\binom{k-j}{2}\right]b}} \frac{1}{(a(k-1)+1)^{2j}} \frac{1}{(a(k-1+j)+1)^{k-j}}.$$

We can now write the variance in more detail and in Chebychev's inequality we now have that

$$P[\mathbf{X}=0] \leq \frac{\mathrm{Var}[\mathbf{X}]}{E[\mathbf{X}]^2} = \frac{\sum_{j=0}^{k} \frac{\binom{k}{k-j}\binom{n}{k}\binom{n-k}{j}}{n^{\left[2\binom{k}{2}-\binom{k-j}{2}\right]b}} \frac{1}{(a(k-1)+1)^{2j}} \frac{1}{(a(k-1+j)+1)^{k-j}} - E[\mathbf{X}]^2}{E[\mathbf{X}]^2}.$$

We claim that the above value is going to 0 as $n \to \infty$. We show this by considering the righthand side of the inequality in three parts and showing that each of the parts individually goes to zero or is negative.

First we note that when $j = 0$ each of the $u_i = y_i$, $1 \le i \le k$, and the value is

$$\frac{\binom{n}{k}}{n^{\binom{k}{2}b}} \frac{1}{(a(k-1)+1)^k} = E[\mathbf{X}].$$

Therefore in the Chebychev's we have that as $n \to \infty$

$$\frac{\frac{\binom{n}{k}}{n^{\binom{k}{2}b}} \frac{1}{(a(k-1)+1)^k}}{E[\mathbf{X}]^2} = \frac{E[\mathbf{X}]}{E[\mathbf{X}]^2} \to 0.$$

Next when $j = k$ and none of the terms overlap the value is

$$\frac{\binom{n}{k}\binom{n-k}{k}\frac{1}{(a(k-1)+1)^{2k}}}{n^{2\binom{k}{2}b}} < E[\mathbf{X}]^2.$$

Therefore in the Chebychev's inequality we have that as $n \to \infty$

$$\frac{\frac{\binom{n}{k}\binom{n-k}{k}\frac{1}{(a(k-1)+1)^{2k}}}{n^{2\binom{k}{2}b}} - E[\mathbf{X}]^2}{E[\mathbf{X}]} < 0.$$

So, consider any j, $1 \le j \le k-1$. Then the term in the summation is

$$\frac{\frac{\binom{k}{k-j}\binom{n}{k}\binom{n-k}{j}}{n^{\left[2\binom{k}{2}-\binom{k-j}{2}\right]b}} \frac{1}{(a(k-1)+1)^{2j}} \frac{1}{(a(k-1+j)+1)^{k-j}}}{E[\mathbf{X}]^2} = \frac{\frac{\binom{k}{k-j}\binom{n}{k}\binom{n-k}{j}}{n^{\left[2\binom{k}{2}-\binom{k-j}{2}\right]b}} \frac{1}{(a(k-1)+1)^{2j}} \frac{1}{(a(k-1+j)+1)^{k-j}}}{\frac{\binom{n}{k}^2}{n^{2\binom{k}{2}b}} \frac{1}{(a(k-1)+1)^k}}$$

which for fixed k is

$$\asymp \frac{n^k n^j}{n^{[2\binom{k}{2}-\binom{k-j}{2}]b}} \frac{n^{2\binom{k}{2}b}}{n^{2k}} \frac{(a(k-1)+1)^{k-2j}}{(a(k-1+j)+1)^{k-j}}$$

$$= C(a,k,b,j)/n^{k-j-\binom{k-j}{2}b}$$

and is simply a constant, that does not depend on n, times $1/n^{k-j-\binom{k-j}{2}b}$. Whenever $b < \frac{2}{k-1}$ we have that the power on n is

$$k-j-\binom{k-j}{2}b > k-j-\binom{k-j}{2}\frac{2}{k-1} = \frac{j(k-j)}{k-1} \geq 1$$

since $1 \leq j \leq k-1$. Therefore whenever $1 \leq j \leq k-1$ the term in Chebychev's inequality is

$$\frac{\frac{\binom{k}{k-j}\binom{n}{k}\binom{n-k}{j}}{n^{\left[2\binom{k}{2}-\binom{k-j}{2}\right]b}} \frac{1}{(a(k-1)+1)^{2j}} \frac{1}{(a(k-1+j)+1)^{k-j}}}{E[\mathbf{X}]^2} < \frac{1}{n} \to 0$$

as $n \to \infty$.

Hence, $P[\mathbf{X}=0] \to 0$ as $n \to \infty$ and we will almost surely have a clique of size k whenever $b < \frac{2}{k-1}$.

QED.

So, the threshold for the appearance of a clique of size k is $b = \frac{2}{k-1}$, i.e., when the probability mapping is $f(r) = \frac{r}{n^{2/(k-1)}}$ which agrees with the threshold function for the appearance of cliques of size k in a Erdős-Rényi random graph. Therefore when $b < \frac{2}{k-1}$ almost every graph will have a clique of size k and whenever $b > \frac{2}{k-1}$ almost no graph will have a clique of size k, which gives us the following result.

Corollary 3.1.6. *Let G be drawn from $\mathcal{DS}[n, a, b, 1]$ with $b \in (0, 2)$. Let $A = \{\frac{2}{i-1} : i \in \mathbb{Z}_+\}$. Let $k_b = \max\{k : b < \frac{2}{k-1}, k \in \mathbb{Z}_+\}$. Then the clique number of G satisfies*

$$\omega(G) \in \begin{cases} \{k_b, k_b + 1\} & b \in A \\ \\ \{k_b\} & b \notin A \end{cases}.$$

Proof:

First note that since $b < \frac{2}{k_b - 1}$, by Theorem 3.1.5, G will almost surely have a clique of size k_b and so $\omega(G) \geq k_b$.

Now, if $b \in A$, then $b = \frac{2}{k_b} = \frac{2}{(k_b+1)-1}$. Therefore, $\forall k > k_b + 1$, $b > \frac{2}{k-1}$. Hence by Theorem 3.1.5, G will almost surely not contain a clique of size k. Therefore, $\omega(G) \leq k_b + 1$.

Similarly, if $b \notin A$, then $b > \frac{2}{k_b} = \frac{2}{(k_b+1)-1}$ and so again by Theorem 3.1.5, G will almost surely not contain a clique of size $k_b + 1$. Hence, $\omega(G) \leq k_b$.

QED.

3.1.2 Results when $b \in (0, 1)$

In this section we consider the case when the probability mapping is $f(r) = \frac{r}{n^b}$ for $b \in (0, 1)$. Assume that n is large. We study the degree distribution of G.

Proposition 3.1.7. *Let G be drawn from $\mathcal{DS}[n, a, 1]$ with $b \in (0, 1)$. The expected number of vertices of degree zero in G is*

$$E[|\{v : d(v) = 0\}|] \sim \left(\frac{1}{a} (1 + a)^{\frac{1}{a}} \Gamma \left(\frac{1}{a} \right) \right) n^{\frac{a-1+b}{a}}.$$

Proof:

Choose $v \in V(G)$ fixed, but arbitrary, and denote the vector of v by y. Let $x_1, x_2, \ldots, x_{n-1}$ be the vectors of the remaining $n - 1$ vertices in $V(G) - \{v\}$. Then

$$P[d(v) = 0] = \int_0^1 \int_0^1 \cdots \int_0^1 (1 - \frac{x_1 y}{n^b}) \cdots (1 - \frac{x_{n-1} y}{n^b})$$
$$\cdot\, g(x_1) \,\cdots\, g(x_{n-1})\, g(y)\, dx_1 \cdots dx_{n-1}\, dy$$

which is separable. Noting that for each of the x_i's

$$\int_0^1 (1 - \frac{x_i y}{n^b}) g(x_i) dx_i = \int_0^1 \frac{1}{a} (1 - \frac{x_i y}{n^b}) x_i^{\frac{1-a}{a}}\, dx_i = 1 - \frac{y}{n^b(a + 1)}$$

we have that

$$P[d(v) = 0] = \int_0^1 \left(1 - \frac{y}{n^b(a + 1)} \right)^{n-1} \frac{1}{a}\, y^{\frac{1-a}{a}}\, dy.$$

Substituting $\hat{y} = y^{1/a}$ for y in the integral we have that

$$P[d(v) = 0] = \int_0^1 \left(1 - \frac{\hat{y}^a}{n^b(a+1)}\right)^{n-1} d\hat{y}.$$

From $\hat{y} = t/n^{\frac{1-b}{a}}$, one has $d\hat{y} = dt/n^{\frac{1-b}{a}}$ and $0 \leq \hat{y} \leq 1$ yields $0 \leq t \leq n^{\frac{1-b}{a}}$. Therefore, substituting again and noting that $b < 1$, we have that

$$P[d(v) = 0] = \int_0^{n^{\frac{1-b}{a}}} \left(1 - \frac{t^a}{(a+1)n}\right)^{n-1} \frac{dt}{n^{\frac{1-b}{a}}}$$

$$\sim \int_0^{\infty} \exp\left\{\frac{-t^a}{a+1}\right\} \frac{dt}{n^{\frac{1-b}{a}}} = \left(\frac{1}{a}(1+a)^{\frac{1}{a}}\,\Gamma\left(\frac{1}{a}\right)\right) \frac{1}{n^{\frac{1-b}{a}}}.$$

Therefore the expected number of vertices of degree zero is

$$E[|\{v : d(v) = 0\}|] = n\,P[d(v) = 0] \sim \left(\frac{1}{a}(1+a)^{\frac{1}{a}}\,\Gamma\left(\frac{1}{a}\right)\right) n^{\frac{a-1+b}{a}}.$$

QED.

We see that the expected number of isolated vertices in G is asymptotic to a constant times $n^{\frac{a-1+b}{a}}$. The above result can be generalized to vertices of degree k where $k \ll n$ as follows:

Proposition 3.1.8. *Let k be a fixed nonnegative integer. Let $\lambda(k)$ be the number of vertices of degree k in a random dot product graph drawn from $\mathcal{D}[n, a, b, 1]$ with*

$b \in (0,1)$. *Then*

$$E[\lambda(k)] \sim \left(\frac{1}{k!\,a}(1+a)^{\frac{1}{a}}\Gamma\left(\frac{1}{a}+k\right)\right) n^{\frac{a-1+b}{a}}$$

as $n \to \infty$.

In the dense version of the Random Dot Product Graph we discussed the idea of clustering in Section 2.3.2. This property clearly distinguishes the dense version of the model from an Erdős-Rényi random graph which does not exhibit any clustering. We have an analogous result for the Sparse Random Dot Product Graph.

Proposition 3.1.9. *Let* G *be drawn from* $\mathcal{DS}[n,a,b,1]$ *with* $b \in (0,1)$. *Let* $v \in V(G)$ *and let* $N(v)$ *be the open neighborhood of* v. *Then* $\forall\, u, w \in N(v)$,

$$P[u \sim w | u \sim v \sim w] = \frac{(a+1)^2}{n^b(2a+1)^2}.$$

Note the proof is almost identical to that of Lemma 2.3.6 and is therefore omitted.

While it is true that for any vertices u, v, and w, the $P[u \sim v] < P[u \sim v | u \sim w \sim v]$ and so some clustering does exist, we see that $P[u \sim v | u \sim w \sim v] - P[u \sim v] = \frac{(1+a)^2}{n^b(1+2a)^2} - \frac{1}{n^b(a+1)^2} \to 0$ as $n \to \infty$. So, in the Sparse Random Dot Product Graph we cannot use clustering as an example of a limiting behavior different from that of an Erdős-Rényi random graph.

It is well known that $\frac{\log n}{n}$ is a threshold function for the disappearance of isolated vertices in an Erdős-Rényi random graph. We will show that in the Sparse Random Dot Product Graph no such threshold exists.

Proposition 3.1.10. *Let G be drawn from $\mathcal{DS}[n, a, b, 1]$ with $b \in (0, 1)$. For each vertex $v \in V(G)$, let $\mathbf{X}_v$ be the indicator that v is isolated. Let $\mathbf{X} = \sum_{v \in V(G)} \mathbf{X}_v$ be the number of isolated vertices in G. Then $P[\mathbf{X} > \mathbf{0}] \to 1$ as $n \to \infty$ and almost every graph has isolated vertices.*

Proof:

First recall from Proposition 3.1.7 that $E[\mathbf{X}] \sim \left(\frac{1}{a}(1+a)^{\frac{1}{a}}\Gamma\left(\frac{1}{a}\right)\right) n^{\frac{a-1+b}{a}} \to \infty$ for all $b \in (0, 1)$. Also, by Chebychev's inequality we know that $P[\mathbf{X} = 0] \leq \mathrm{Var}(\mathbf{X})/E[\mathbf{X}]^2 = \frac{E[\mathbf{X}^2] - E[\mathbf{X}]^2}{E[\mathbf{X}]^2}$ and so we calculate $E[\mathbf{X}^2]$.

$$E[\mathbf{X}^2] = E\left[\left(\sum_{v \in V(G)} \mathbf{X}_v\right)^2\right]$$

(where $\mathbf{X}_v$ is the indicator that v is isolated)

$$= E\left[\left(\sum_{v \in V(G)} \mathbf{X}_v\right)\left(\sum_{u \in V(G)} \mathbf{X}_u\right)\right]$$

$$= \sum_{u,v \in V(G)} E[\mathbf{X}_u \mathbf{X}_v].$$

If $u = v$ then $\mathbf{X}_u \mathbf{X}_v = \mathbf{X}_v$ and so we have that

$$E[\mathbf{X}^2] = \sum_{v \in V(G)} E[\mathbf{X}_v] + \sum_{u,v \in V(G): u \neq v} E[\mathbf{X}_u \mathbf{X}_v]$$

$$= E[\mathbf{X}] + 2\binom{n}{2} P[u \text{ and } v \text{ are both isolated}].$$

Now, $u, v \in V(G)$ with vectors x and y, respectively. Let $x_1, x_2, \ldots, x_{n-2}$ be the vectors of the remaining $n-2$ vertices in $V(G) - \{u, v\}$. Then

$$P[u \text{ and } v \text{ are both isolated}] =$$

$$= \int_0^1 \int_0^1 \cdots \int_0^1 (1 - \frac{xy}{n^b}) \left((1 - \frac{x_1 x}{n^b})(1 - \frac{x_1 y}{n^b})\right) \cdots \left((1 - \frac{x_{n-2} x}{n^b})(1 - \frac{x_{n-2} y}{n^b})\right)$$

$$\cdot g(x_1) \cdots g(x_{n-1})\, g(y)\, dx_1 \cdots dx_{n-2}\, dx\, dy$$

which is separable. Noting that for each of the x_i's

$$\int_0^1 (1 - \frac{x_i x}{n^b})(1 - \frac{x_i y}{n^b}) g(x_i) dx_i = 1 - \frac{x}{n^b(a+1)} - \frac{y}{n^b(a+1)} + \frac{xy}{n^{2b}(2a+1)}$$

we have that

$$P[u \text{ and } v \text{ are both isolated}] =$$

$$= \int_0^1 \int_0^1 (1 - \frac{xy}{n^b}) \left(1 - \frac{x}{n^b(a+1)} - \frac{y}{n^b(a+1)} + \frac{xy}{n^{2b}(2a+1)}\right)^{n-2} \frac{1}{a}\, x^{\frac{1-a}{a}} \frac{1}{a}\, y^{\frac{1-a}{a}}\, dx\, dy$$

and substituting $\hat{x} = x^{1/a}$ and $\hat{y} = y^{1/a}$ for y in the integral we have that

$$= \int_0^1 \int_0^1 (1 - \frac{\hat{x}^a \hat{y}^a}{n^b}) \left(1 - \frac{\hat{x}^a}{n^b(a+1)} - \frac{\hat{y}^a}{n^b(a+1)} + \frac{\hat{x}^a \hat{y}^a}{n^{2b}(2a+1)}\right)^{n-2} d\hat{x}\, d\hat{y}$$

$$\leq \int_0^1 \int_0^1 \left(1 - \frac{\hat{x}^a}{n^b(a+1)} - \frac{\hat{y}^a}{n^b(a+1)} + \frac{\hat{x}^a \hat{y}^a}{n^{2b}(2a+1)}\right)^{n-2} d\hat{x}\, d\hat{y}$$

(since $(1 - \frac{\hat{x}^a \hat{y}^a}{n^b}) \leq 1$)

$$\leq \int_0^1 \int_0^1 \exp\left[-(n-2)\left(\frac{\hat{x}^a}{n^b(a+1)} + \frac{\hat{y}^a}{n^b(a+1)} - \frac{\hat{x}^a \hat{y}^a}{n^{2b}(2a+1)}\right)\right] d\hat{x}\, d\hat{y}$$

$$= \int_0^1 \exp\left[-(n-2)\left(\frac{\hat{x}^a}{n^b(a+1)}\right)\right]$$

$$\cdot \int_0^1 \exp\left[-(n-2)\left(\frac{\hat{y}^a}{n^b(a+1)} - \frac{\hat{x}^a \hat{y}^a}{n^{2b}(2a+1)}\right)\right] d\hat{x}\, d\hat{y}$$

$$\leq \int_0^1 \exp\left[-(n-2)\left(\frac{\hat{x}^a}{n^b(a+1)}\right)\right]$$

$$\cdot \int_0^\infty \exp\left[-(n-2)\left(\frac{\hat{y}^a}{n^b(a+1)} - \frac{\hat{x}^a \hat{y}^a}{n^{2b}(2a+1)}\right)\right] d\hat{x}\, d\hat{y}$$

$$= \int_0^1 \exp\left[-(n-2)\left(\frac{\hat{x}^a}{n^b(a+1)}\right)\right]$$

$$\cdot \left(\frac{1}{a}\,\Gamma(\frac{1}{a})\left[(n-2)\left(\frac{1}{n^b(a+1)} - \frac{\hat{x}^a}{n^{2b}(2a+1)}\right)\right]^{-\frac{1}{a}}\right) d\hat{x}$$

$$\leq \int_0^1 \exp\left[-(n-2)\left(\frac{\hat{x}^a}{n^b(a+1)}\right)\right]$$

$$\cdot \left(\frac{1}{a}\,\Gamma(\frac{1}{a})\left[(n-2)\left(\frac{1}{n^b(a+1)} - \frac{1}{n^{2b}(2a+1)}\right)\right]^{-\frac{1}{a}}\right) d\hat{x}$$

(since $\hat{x}^a \leq 1$)

$$\leq \left(\frac{1}{a}\,\Gamma(\frac{1}{a})\left[(n-2)\left(\frac{1}{n^b(a+1)}-\frac{1}{n^{2b}(2a+1)}\right)\right]^{-\frac{1}{a}}\right)$$

$$\cdot\int_0^\infty \exp\left[-(n-2)\left(\frac{\hat{x}^a}{n^b(a+1)}\right)\right]d\hat{x}$$

$$= \left(\frac{1}{a}\,\Gamma(\frac{1}{a})\left[(n-2)\left(\frac{1}{n^b(a+1)}-\frac{1}{n^{2b}(2a+1)}\right)\right]^{-\frac{1}{a}}\right)$$

$$\cdot\left(\frac{1}{a}\,\Gamma(\frac{1}{a})\left[(n-2)\left(\frac{1}{n^b(a+1)}\right)\right]^{-\frac{1}{a}}\right)$$

$$= \left(\frac{1}{a}\,\Gamma(\frac{1}{a})\right)^2(n-2)^{-\frac{2}{a}}\left[\left(\frac{1}{n^b(a+1)}-\frac{1}{n^{2b}(2a+1)}\right)\left(\frac{1}{n^b(a+1)}\right)\right]^{-\frac{1}{a}}.$$

So, we have that

$$E[\mathbf{X}^2] \leq E[\mathbf{X}]$$

$$+2\binom{n}{2}\left(\frac{1}{a}\,\Gamma(\frac{1}{a})\right)^2(n-2)^{-\frac{2}{a}}\left[\left(\frac{1}{n^b(a+1)}-\frac{1}{n^{2b}(2a+1)}\right)\left(\frac{1}{n^b(a+1)}\right)\right]^{-\frac{1}{a}}.$$

Hence

$$P[\mathbf{X}=0] \leq \frac{E[\mathbf{X}^2]-E[\mathbf{X}]^2}{E[\mathbf{X}]^2}$$

$$\leq \frac{E[\mathbf{X}]+2\binom{n}{2}\left(\frac{1}{a}\,\Gamma(\frac{1}{a})\right)^2(n-2)^{-\frac{2}{a}}\left[\left(\frac{1}{n^b(a+1)}-\frac{1}{n^{2b}(2a+1)}\right)\left(\frac{1}{n^b(a+1)}\right)\right]^{-\frac{1}{a}}-E[\mathbf{X}]^2}{E[\mathbf{X}]^2}$$

$$\sim \frac{E[\mathbf{X}]}{E[\mathbf{X}]^2}+\frac{\left(\frac{1}{a}\,\Gamma(\frac{1}{a})\right)^2 n^{2-\frac{2}{a}}\left[\left(\frac{1}{n^b(a+1)}-\frac{1}{n^{2b}(2a+1)}\right)\left(\frac{1}{n^b(a+1)}\right)\right]^{-\frac{1}{a}}}{E[\mathbf{X}]^2}-\frac{E[\mathbf{X}]^2}{E[\mathbf{X}]^2}.$$

Now, $\frac{E[\mathbf{X}]}{E[\mathbf{X}]^2} \to 0$ as $n \to \infty$ and $\frac{E[\mathbf{X}]^2}{E[\mathbf{X}]^2} = 1$. Also,

$$\frac{\left(\frac{1}{a}\,\Gamma(\frac{1}{a})\right)^2 n^{2-\frac{2}{a}} \left[\left(\frac{1}{n^b(a+1)} - \frac{1}{n^{2b}(2a+1)}\right)\left(\frac{1}{n^b(a+1)}\right)\right]^{-\frac{1}{a}}}{E[\mathbf{X}]^2} \sim$$

$$\sim \frac{\left(\frac{1}{a}\,\Gamma(\frac{1}{a})\right)^2 n^{2-\frac{2}{a}} \left[\left(\frac{1}{n^b(a+1)} - \frac{1}{n^{2b}(2a+1)}\right)\left(\frac{1}{n^b(a+1)}\right)\right]^{-\frac{1}{a}}}{\left(\left(\frac{1}{a}\,(1+a)^{\frac{1}{a}}\,\Gamma\left(\frac{1}{a}\right)\right) n^{1-\frac{1}{a}+\frac{b}{a}}\right)^2}$$

$$= \frac{\left(\frac{1}{n^b(a+1)} - \frac{1}{n^{2b}(2a+1)}\right)^{-\frac{1}{a}}}{(1+a)^{\frac{1}{a}} n^{\frac{b}{a}}}$$

$$= \frac{\left(\frac{n^{2b}(a+1)(2a+1)}{n^b(2a+1)-(a+1)}\right)^{\frac{1}{a}}}{(1+a)^{\frac{1}{a}} n^{\frac{b}{a}}}$$

$$= \left(\frac{n^b(2a+1)}{n^b(2a+1)-(a+1)}\right)^{\frac{1}{a}}$$

$$\to 1$$

as $n \to \infty$. Therefore

$$P[\mathbf{X}=0] \to 0 + 1 - 1 = 0$$

as $n \to \infty$ and almost every graph contains isolated vertices.

QED.

3.1.3 The Land of Trees: $b \in (1, 2)$

In this section we discuss the behavior of the Sparse Random Dot Product Graph when the parameter $b \in (1, 2)$. With this restriction on b, the graph will, with high probability, contain no cycles and therefore be a forest on trees of size $\frac{b}{b-1}$ or less.

For any graph H, let $P_{\geq}[H]$ be the probability of H appearing as a (not necessarily induced) subgraph of a Sparse Random Dot Product Graph G on a specific set of vertices in a specific order. We begin by showing a general result relating the number of edges in a graph H to $P_{\geq}[H]$.

Proposition 3.1.11. *Let G be drawn from $\mathcal{DS}[n, a, b, 1]$ with $b \in (0, \infty)$. Let H be a graph on $k > 1$ vertices with $m \geq 1$ edges. For any distinct $u_1, u_2, \ldots, u_k \in V(G)$ we have that the probability of H appearing with edges $e_1, e_2, \ldots e_m$ on $u_1, u_2, \ldots, u_k$ is*

$$P_{\geq}[H \text{ on } u_1, u_2, \ldots, u_k] = \frac{C(H)}{n^{m\,b}}$$

where $C(H)$ is a constant that depends on the graph H and does not vary with n or b.

Proof:

Let H be a graph on the k vertices $u_1, u_2, \ldots, u_k$ with edges $e_1, e_2, \ldots, e_m$. Let us assume that each $e_i \in E(H)$ has endpoints $l(e_i)$ and $r(e_i)$ with vectors $x(e_i)$ and $y(e_i)$, respectively, where $l(e_i), r(e_i) \in \{u_1, u_2, \ldots, u_k\}$ and $x(e_i), y(e_i) \in \{x_{u_1}, x_{u_2}, \ldots, x_{u_k}\}$. Also for each $1 \leq i \leq k$, let r_i be the number of edges incident to u_i. Then

$$P_{\geq}[H \text{ on } u_1, u_2, \ldots, u_k] = \int_0^1 \cdots \int_0^1 \frac{x(e_1)\, y(e_1)}{n^b} \frac{x(e_2)\, y(e_2)}{n^b} \cdots \frac{x(e_m)\, y(e_m)}{n^b}$$

$$g(x_{u_1})\, g(x_{u_2}) \cdots g(x_{u_k})\, dx_{u_1}\, dx_{u_2} \cdots dx_{u_k}$$

$$= \frac{1}{n^{m\,b}} \int_0^1 \cdots \int_0^1 x_{u_1}^{r_1}\, x_{u_2}^{r_2} \cdots x_{u_k}^{r_k} g(x_{u_1})\, g(x_{u_2}) \cdots g(x_{u_k})\, dx_{u_1}\, dx_{u_2} \cdots dx_{u_k}$$

$$= \frac{1}{n^{m\,b}} \left(\int_0^1 x_{u_1}^{r_1} g(x_{u_1})\, dx_{u_1} \right) \left(\int_0^1 x_{u_2}^{r_2} g(x_{u_2})\, dx_{u_2} \right) \cdots \left(\int_0^1 x_{u_k}^{r_k} g(x_{u_k})\, dx_{u_k} \right)$$

$$= \frac{1}{n^{m\,b}} \left(\frac{1}{ar_1 + 1} \right) \left(\frac{1}{ar_2 + 1} \right) \cdots \left(\frac{1}{ar_k + 1} \right)$$

$$= \frac{C(H)}{n^{m\,b}}$$

where $C(H) = \left(\frac{1}{ar_1+1}\right)\left(\frac{1}{ar_2+1}\right) \cdots \left(\frac{1}{ar_k+1}\right)$ and does not vary with n or b. One should note that this probability is essentially a constant over n raised to the number of edges in the graph times b.

QED.

So, for trees we have the following result.

Corollary 3.1.12. *Let G be drawn from $\mathcal{DS}[n, a, b, 1]$ with $b \in (0, \infty)$. Let $k \in \mathbb{N}, k > 1$. For any distinct $u_1, u_2, \ldots, u_k \in V(G)$ we have that the probability of T appearing with $k-1$ edges $e_1, e_2, \ldots e_m$ on $u_1, u_2, \ldots, u_k$ is*

$$P_{\geq}[T \text{ on } u_1, u_2, \ldots, u_k] = \frac{C(H)}{n^{(k-1)\,b}}$$

where $C(T)$ is a constant that depends on the tree T and does not vary with n or b.

Therefore the expected number of copies of a tree T is $E[T] \asymp \frac{\binom{n}{k}}{n^{(k-1)b}} \asymp n^{k-(k-1)b}$ and so if $b > \frac{k}{k-1}$ then we do not expect to see the tree T.

Theorem 3.1.13. *Let G be drawn from $\mathcal{DS}[n, a, b, 1]$ with $b \in (0, \infty)$. Let $k \in \mathbb{Z}_+$. A threshold for the appearance of a tree on k vertices is $b = \frac{k}{k-1}$.*

Proof:

Let T be a tree on k vertices. Let the random variable $\mathbf{X}$ be the number of copies of T in G. Then we know by Corollary 3.1.12 that $E(\mathbf{X}) \asymp n^{k-(k-1)b}$. Let $b = \frac{k}{k-1}$ be our candidate for the threshold for the appearance of T. Now, by Markov's inequality

$$P[\mathbf{X} \geq 1] \leq E[\mathbf{X}] \asymp n^{k-(k-1)b}.$$

So whenever $b > \frac{k}{k-1}$, we have that $E[\mathbf{X}] \to 0$ and therefore the $P[\mathbf{X} = 0] \to 1$ as $n \to \infty$ and almost no graph contains a copy of T.

Also, whenever, $b < \frac{k}{k-1}$, then $E[\mathbf{X}] \to \infty$ and we apply the second moment method. By Chebychev's inequality we know that $P[\mathbf{X} = 0] \leq \mathrm{Var}(\mathbf{X})/E[\mathbf{X}]^2$.

Let $\mathcal{T}$ be the set of all copies of T possible on $V(G)$. For any $T' \in \mathcal{T}$, let $I_{T'}$ be the indicator function for the specific copy T', then $\mathbf{X} = \sum_{T' \in \mathcal{T}} I_{T'}$. The variance of $\mathbf{X}$ is

$$\mathrm{Var}[\mathbf{X}] = E[\mathbf{X}^2] - E[\mathbf{X}]^2 \asymp E[\mathbf{X}^2] - \left[n^{k-(k-1)b}\right]^2$$

and so we calculate $E[\mathbf{X}^2]$.

$$E[\mathbf{X}^2] = E\left[\left(\sum_{T'\in\mathcal{T}} I_{T'}\right)^2\right]$$

$$= E\left[\left(\sum_{T'\in\mathcal{T}} I_{T'}\right)\left(\sum_{T''\in\mathcal{T}} I_{T''}\right)\right]$$

$$= \sum_{T',T''\in\mathcal{T}} E[I_{T'}I_{T''}].$$

Now in each term of the summation T' and T'' share up to $k-1$ edges and we can rewrite the summation as

$$E[\mathbf{X}^2] = \sum_{i=0}^{k-1} \sum_{\substack{T',T''\in\mathcal{T} \\ |E(T')\cap E(T'')|=i}} E[I_{T'}I_{T''}]$$

$$= \sum_{i=0}^{k-1} \sum_{\substack{T',T''\in\mathcal{T} \\ |E(T')\cap E(T'')|=i}} E[I_{T'}I_{T''}]$$

$$= \sum_{i=0}^{k-1} \sum_{\substack{T',T''\in\mathcal{T} \\ |E(T')\cap E(T'')|=i}} P_{\geq}[T'\cup T''].$$

Now, by a discussion similar to that in the proof of Proposition 3.1.12 it can be seen that

$$P_{\geq}[T'\cup T''] = \frac{C(T',T'')}{n^{(|E(T')\cup E(T'')|)b}} = \frac{C(T',T'')}{n^{(2(k-1)-i)b}}$$

where i is the number of edges that T and T'' share and $C(T',T'')$ is a constant, not dependent on n or b. Hence, we have that

$$E[\mathbf{X}^2] = \sum_{i=0}^{k-1} \sum_{\substack{T',T''\in\mathcal{T} \\ |E(T')\cap E(T'')|=i}} \frac{C(T',T'')}{n^{(2(k-1)-i)b}}$$

$$= \sum_{\substack{T',T''\in\mathcal{T} \\ |E(T')\cap E(T'')|=0}} \frac{C(T',T'')}{n^{2(k-1)b}} + \sum_{i=1}^{k-1} \sum_{\substack{T',T''\in\mathcal{T} \\ |E(T')\cap E(T'')|=i}} \frac{C(T',T'')}{n^{(2(k-1)-i)b}}.$$

If T and T'' share $i > 1$ edges, then they also share at least $i+1$ vertices and so for each $i > 1$ we have at no more than $\binom{n}{2k-(i+1)}c_T$ possible choices for the pair (T',T''), where c_T depends only on T and is a constant with respect to n and b. So we have that

$$E[\mathbf{X}^2] \leq \sum_{\substack{T',T''\in\mathcal{T} \\ |E(T')\cap E(T'')|=0}} \frac{C(T',T'')}{n^{2(k-1)b}} + \sum_{i=1}^{k-1} \binom{n}{2k-(i+1)} c_T \frac{C(T',T'')}{n^{(2(k-1)-i)b}}.$$

Now, returning to Chebychev's inequality we have that

$$P[\mathbf{X}=0] \leq \frac{\mathrm{Var}(\mathbf{X})}{E[\mathbf{X}]^2} = \frac{E[\mathbf{X}^2]-E[\mathbf{X}]^2}{E[\mathbf{X}]^2}$$

$$\leq \frac{\sum_{\substack{T',T''\in\mathcal{T} \\ |E(T')\cap E(T'')|=0}} \frac{C(T',T'')}{n^{2(k-1)b}} + \sum_{i=1}^{k-1} \binom{n}{2k-(i+1)} c_T \frac{C(T',T'')}{n^{(2(k-1)-i)b}} - E[\mathbf{X}]^2}{E[\mathbf{X}]^2}$$

$$\asymp \left[\sum_{\substack{T',T''\in\mathcal{T} \\ |E(T')\cap E(T'')|=0}} \frac{1}{n^{2(k-1)b}} + \sum_{i=1}^{k-1} \frac{n^{2k-(i+1)}}{n^{(2(k-1)-i)b}} - \left(n^{k-(k-1)b}\right)^2 \right] \frac{1}{\left(n^{k-(k-1)b}\right)^2}$$

$$= \frac{\left(\sum_{\substack{T',T''\in\mathcal{T}\\|E(T')\cap E(T'')|=0}} \frac{1}{n^{2(k-1)b}}\right) - \left(n^{k-(k-1)b}\right)^2}{\left(n^{k-(k-1)b}\right)^2} + \frac{\sum_{i=1}^{k-1} \frac{n^{2k-(i+1)}}{n^{(2(k-1)-i)b}}}{\left(n^{k-(k-1)b}\right)^2}.$$

We will examine the first term of the summation separately from the remaining terms. In the first term, any pair, T' and T'', share no edges, but may share vertices. Let us assume that they share $j \in \{0, k\}$ vertices, then the term can be written as

$$\frac{\left(\sum_{j=0}^{k} \sum_{\substack{T',T''\in\mathcal{T}\\|E(T')\cap E(T'')|=0\\|V(T')\cap V(T'')|=j}} \frac{1}{n^{2(k-1)b}}\right) - \left(n^{k-(k-1)b}\right)^2}{\left(n^{k-(k-1)b}\right)^2}$$

(and is less than)

$$\leq \frac{\left(\sum_{j=0}^{k} \frac{\binom{n}{k}\binom{n-k}{k-j}2k!}{n^{2(k-1)b}}\right) - \left(n^{k-(k-1)b}\right)^2}{\left(n^{k-(k-1)b}\right)^2}$$

(which for large n is)

$$\leq \frac{\left(\sum_{j=0}^{k} \frac{\binom{n}{k}\binom{n-k}{k}2k!}{n^{2(k-1)b}}\right) - \left(n^{k-(k-1)b}\right)^2}{\left(n^{k-(k-1)b}\right)^2}$$

$$= \frac{\left((k+1)\frac{\binom{n}{k}\binom{n-k}{k}2k!}{n^{2(k-1)b}}\right) - \left(n^{k-(k-1)b}\right)^2}{\left(n^{k-(k-1)b}\right)^2}$$

$$\asymp \frac{\frac{n^k(n-k)^k}{n^{2(k-1)b}} - \frac{n^k n^k}{n^{2(k-1)b}}}{\left(n^{k-(k-1)b}\right)^2}$$

$$< 0.$$

So, the first term in the summation is less than zero and we only need to consider the remaining terms and so we have that

$$P[\mathbf{X}=0] \asymp \frac{\left(\sum_{\substack{T',T''\in\mathcal{T}\\|E(T')\cap E(T'')|=0}} \frac{1}{n^{2(k-1)b}}\right) - \left(n^{k-(k-1)b}\right)^2}{\left(n^{k-(k-1)b}\right)^2} + \frac{\sum_{i=1}^{k-1} \frac{n^{2k-(i+1)}}{n^{(2(k-1)-i)b}}}{\left(n^{k-(k-1)b}\right)^2}$$

$$< 0 + \sum_{i=1}^{k-1} \frac{\frac{n^{2k-(i+1)}}{n^{(2(k-1)-i)b}}}{\left(n^{k-(k-1)b}\right)^2} = \sum_{i=1}^{k-1} n^{-(i+1)-i\,b}.$$

Now, if for each $i \in \{1, k-1\}$ we can show that $b < \frac{i+1}{i}$, then $P[\mathbf{X}=0] \to 0$ as $n \to \infty$. We know that each $i \le k-1$ and $b < \frac{k}{k-1}$. Additionally, $\frac{i+1}{i}$ is minimized when $i = k-1$. Thus for all $i \in \{1, k-1\}$, $\frac{i+1}{i} \ge \frac{k}{k-1} > b$. Hence, $P[\mathbf{X}=0] \to 0$ as $n \to \infty$ and we will almost surely have a copy of T whenever $b < \frac{k}{k-1}$.

QED.

So, the threshold for the appearance of a tree on k vertices is $b = \frac{k}{k-1}$. Therefore when $b < \frac{k}{k-1}$ almost every graph will have a tree of size k and whenever $b > \frac{2}{k-1}$ almost no graph will have a clique of size k, which gives us the following result.

Corollary 3.1.14. *Let G be drawn from $\mathcal{DS}[n, a, b, 1]$ with $b \in (0, \infty)$. Let k be a fixed positive integer, then whenever*

- $b \in (0, 1]$, *G will almost surely contain all trees on k or fewer vertices,*
- $b \in (1, 2)$, *G will almost surely contain a tree on k vertices if and only if $k < \frac{b}{b-1}$,*
- $b \in (2, \infty)$, *G will almost surely contain no edges.*

Proof:

Theorem 3.1.13 we know that with high probability G will contain a tree on k vertices whenever $b < \frac{k}{k-1}$. Secondly, for all $k \in \mathbf{Z}_+$, $\frac{k}{k-1} > 1$ and thus G will almost surely contain a tree of size k. Also with high probability, $b < \frac{k}{k-1}$ if and only if $k < \frac{b}{b-1}$. So, for all $(2, \infty)$, the only positive integer less than $\frac{b}{b-1}$ is 1, and hence there are no trees on 2 or more vertices. Finally, when $b \in (1, 2)$, with high probability G will contain a tree on k vertices if and only if $k < \frac{b}{b-1}$.

QED.

Thus far we have shown that whenever $b \in (1, 2)$, G will almost surely contain trees on at most $k_b = \lceil \frac{b}{b-1} \rceil - 1$ vertices. We now show that G will almost surely be without cycles and thus a forest.

Proposition 3.1.15. *Let G be drawn from $\mathcal{DS}[n, a, b, 1]$ with $b \in (0, \infty)$. Let $k \in \mathbb{N}, k > 1$. For any distinct $u_1, u_2, \ldots, u_k \in V(G)$ we have that the probability of the k-cycle $u_1, u_2, \ldots, u_k, u_1$ is*

$$P_{\geq}[\textit{the } k\textit{-cycle } u_1, u_2, \ldots, u_k, u_1] = \frac{(2a+1)^{-k}}{n^{kb}}.$$

Proof:

$$P_{\geq}[\text{the } k\text{-cycle } u_1, u_2, \ldots, u_k, u_1] = \int_0^1 \cdots \int_0^1 \frac{x_{u_1}\, x_{u_2}}{n^b} \frac{x_{u_2}\, x_{u_3}}{n^b} \cdots \frac{x_{u_{k-1}}\, x_{u_k}}{n^b} \frac{x_{u_k}\, x_{u_1}}{n^b}$$

$$g(x_{u_1})\, g(x_{u_2}) \cdots g(x_{u_k})\, dx_{u_1}\, dx_{u_2} \cdots dx_{u_k}$$

$$= \frac{1}{n^{kb}} \int_0^1 \cdots \int_0^1 x_{u_1}^2\, x_{u_2}^2 \cdots x_{u_k}^2 g(x_{u_1})\, g(x_{u_2})\, \cdots g(x_{u_k})\, dx_{u_1}\, dx_{u_2} \cdots dx_{u_k}$$

$$= \frac{1}{n^{kb}} \left[\int_0^1 x_{u_i}^2 g(x_{u_i}) dx_{u_i} \right]^k$$

$$= \frac{(2a+1)^{-k}}{n^{kb}}.$$

QED.

Therefore the expected number of of size k-cycles is $\binom{n}{k}(k-1)!\frac{(2a+1)^{-k}}{n^{kb}}$.

Theorem 3.1.16. *Let G be drawn from $\mathcal{DS}[n, a, b, 1]$ with $b \in (1, 2)$. Let $k_b = \lceil \frac{b}{b-1} \rceil - 1$. Then with high probability, G is a forest containing trees with at most k_b vertices.*

Proof:

By Corollary 3.1.14, with high probability G will contain trees on $1, 2, \ldots, k_b$ vertices and will not contain trees on $k_b + 1$ or more vertices. Specifically, G will not contain any path of length $k_b + 1$ or greater. Hence, with high probability, G will not contain a cycle of length $k_b + 1$ or greater. Let $2 \le c \le k_b$, and let the random variable $\mathbf{X}_c$ be the number of cycles of length c in G. Then we know by Proposition 3.1.15 that $E(\mathbf{X}_c) = \binom{n}{c}(c-1)!\frac{(2a+1)^{-c}}{n^{cb}}$. Now, by Markov's inequality

$$P[\mathbf{X}_c \ge 1] \le E[\mathbf{X}_c] \asymp n^{c(1-b)} \le n^{k_b(1-b)} \to 0$$

as $n \to \infty$ since $k_b < \frac{b}{b-1}$. Hence almost no graph contains a cycle of length $c \in$

$\{2, 3, \ldots, k_b\}$. Thus with high probability, G will not contain any cycles and hence will be a forest.

QED.

3.2 Main Results

3.2.1 The Threshold Result

Let H be a graph with k vertices and $l \geq 1$ edges. We define $\varepsilon(H) = \frac{l}{k}$, i.e., the number of edges divided by the number of vertices. Also, let

$$\varepsilon'(H) = \max\{\varepsilon(F) | F \text{ is a nonempty subgraph of } H\}.$$

Then we have the following result about the appearance of any graph H.

Theorem 3.2.1. *Let G be drawn from $\mathcal{DS}[n, a, b, 1]$ with $b \in (0, \infty)$. Let H be a graph with k vertices and $l \geq 1$ edges. Then $b = \frac{1}{\varepsilon'(H)}$ is a threshold for the appearance of H.*

Proof:

Let $\mathcal{H}$ be the set of all graphs that lie on a subset of the vertex set of G and are isomorphic to H. Note that $|\mathcal{H}| = \Theta(n^k)$. Also, $\forall H' \in \mathcal{H}$ let $X_{H'}$ be the indicator function for when H' is a subgraph of G. Let X be number of subgraphs of G that are isomorphic to H. Then $X = \sum_{H' \subseteq \mathcal{H}} X_{H'}$ and by linearity of expectation we have that

$$E[X] = \sum_{H' \subseteq \mathcal{H}} E[X_{H'}] = \Theta(n^k) P_{\geq}[H'].$$

Let $x_1, \ldots, x_k$ be the vectors of vertices $h_1, \ldots, h_k$ of H', respectively. Then

$$P_{\geq}[H'] =$$

$$= \int_0^1 \cdots \int_0^1 \prod_{1 \leq i < j \leq k} \left(\frac{x_i x_j}{n^b} \right)^{I\{h_i \sim h_j \text{ in } H'\}} g(x_1) \cdots g(x_k) dx_1 \cdots dx_k$$

$$= \frac{1}{n^{lb}} \int_0^1 \cdots \int_0^1 \prod_{1 \leq i < j \leq k} (x_i x_j)^{I\{h_i \sim h_j \text{ in } H'\}} g(x_1) \cdots g(x_k) dx_1 \cdots dx_k$$

$$= \frac{1}{n^{lb}} C(H', a)$$

where $C(H', a)$ depends only on H' and a and is a constant with respect to n and b. Also, for any $H', H'' \subseteq \mathcal{H}$, $C(H', a) = C(H'', a) = C(H, a)$. Hence

$$E[X] = \Theta(n^k) \frac{1}{n^{lb}} C(H, a) = \Theta(n^{k-lb}).$$

Let $F = (V(F), E(F))$ be a subgraph of H such that $\varepsilon'(H) = \varepsilon(F)$ and let $b > \frac{1}{\varepsilon'(H)}$. Also, let Y be the number of subgraphs isomorphic to F on G. Then a discussion similar to that above, $E[Y] = \Theta(n^{|V(F)| - |E(F)| b})$. Now, $b > \frac{1}{\varepsilon'(H)}$ gives us that

$$|V(F)| - |E(F)| b < |V(F)| - |E(F)| \frac{1}{\varepsilon'(H)}$$

$$= |V(F)| - |E(F)|\frac{1}{\varepsilon(F)} = |V(F)| - |E(F)|\frac{|V(F)|}{|E(F)|} = 0$$

and so $E[Y] \to 0$. Hence by Markov's inequality $P[Y > 0] \leq E[Y] \to 0$ and almost no graph contains F as a subgraph, therefore since $F \subseteq H$, almost no graph contains H as a subgraph when $b > \frac{1}{\varepsilon'(H)}$.

When $b < \frac{1}{\varepsilon'(H)}$ we see that $k - lb > k - l\frac{1}{\varepsilon'(H)} > k - l\frac{1}{\varepsilon(H)} = k - l\frac{k}{l} = 0$ and hence $E[X] = \Theta(n^{k-lb}) \to \infty$ and we use the second moment method. That is we show that $P[\mathbf{X} = 0] \leq \frac{\mathrm{Var}[\mathbf{X}]}{E[\mathbf{x}]^2} = \frac{E[\mathbf{X}^2] - E[\mathbf{X}]^2}{E[\mathbf{X}]^2} \to 0$.

For any subgraph F of H, consider the set $\mathcal{H}_F^2 = \{(H', H'') \in \mathcal{H}^2 | H' \bigcap H'' = F\}$. Then by a discussion very similar to that above we see that

$$P_{\geq}[H' \bigcup H'' \subseteq G | H' \bigcap H'' = F] = \frac{1}{n^{(2l - |E(F)|)b}} C(H' \bigcup H'', a)$$

since there are $2l - |E(F)|$ edges in $H' \bigcup H''$. Also, $\frac{|E(F)|}{|V(F)|} \leq \varepsilon'(H)$, hence

$$P_{\geq}[H' \bigcup H'' \subseteq G | H' \bigcap H'' = F] \leq \frac{1}{n^{(2l - |V(F)|\varepsilon'(H))b}} C(H' \bigcup H'', a).$$

Let $A_F = \sum_{H' \bigcup H'' = F} P_{\geq}[H' \bigcap H'' \subset G]$, then $E[\mathbf{X}^2] = \sum_F A_F$ and we need to show that $P[\mathbf{X} = 0] \leq \frac{\sum_F A_F - E[\mathbf{X}]^2}{E[\mathbf{X}]^2} \to 0$.

First consider the case where $|V(F_0)| = 0$, i.e., F_0 is the empty graph. Then H'

and H'' appearing in G are independent events and

$$A_{F_0} = \sum_{H' \bigcap H'' = F_0} P_{\geq}[H' \bigcup H'' \subset G] = \sum_{H' \bigcap H'' = F_0} P_{\geq}[H' \subset G] P_{\geq}[H'' \subset G]$$

$$= \sum_{H' \bigcap H'' = F_0} \left(\frac{1}{n^{lb}} C(H', a)\right) \left(\frac{1}{n^{lb}} C(H'', a)\right)$$

$$= \sum_{H' \bigcap H'' = F_0} \left(\frac{1}{n^{lb}} C(H, a)\right)^2$$

$$\leq |\mathcal{H}|^2 \left(\frac{1}{n^{lb}} C(H, a)\right)^2 = E[\mathbf{X}]^2.$$

Hence for the empty graph F_0, $A_{F_0} \leq E[\mathbf{X}]^2$.

Now assume that $|V(F)| > 0$ and let $f = |V(F)|$. Also, let $\sum' = \sum_{H' \in \mathcal{H}}$ and $\sum'' = \sum_{H'' \in \mathcal{H}}$ so that $\sum_{H' \bigcap H'' = F} = \sum' \sum''_{H' \bigcap H'' = F}$. Finally, let h be the number of isomorphic copies of H (or H' or H'' since they are isomorphic to each other) on a fixed set of size k. Then for a fixed H' the inner summation ranges over at most $\binom{k}{f}\binom{n-k}{k-f} h$ terms and so

$$A_F = \sum_{H' \bigcap H'' = F} P_{\geq}[H' \bigcup H'' \subset G]$$

$$= \sum{}' \sum{}''_{H' \bigcap H'' = F} P_{\geq}[H' \bigcup H'' \subset G]$$

$$\leq \sum{}' \binom{k}{f} \binom{n-k}{k-f} h P_{\geq}[H' \bigcup H'' \subset G]$$

$$\leq \sum{}' \binom{k}{f}\binom{n-k}{k-f} h \frac{1}{n^{(2l-f\varepsilon'(H))b}} C(H' \bigcup H'', a).$$

There are $|\mathcal{H}| = \Theta(n^k)$ possible H''s and so we have

$$A_F \leq \Theta(n^k)\binom{k}{f}\binom{n-k}{k-f} h \frac{1}{n^{(2l-f\varepsilon'(H))b}} C(H' \bigcup H'', a)$$

$$= \left(\frac{\Theta(n^k)C^*}{n^l}\right)\left(C^{**}\frac{n^{k-f}}{n^{(2l-f\varepsilon'(H))b}}\right)$$

$$= E[\mathbf{X}]O\left(\frac{n^{k-f}}{n^{(2l-f\varepsilon'(H))b}}\right)$$

where C^* and C^{**} are constants independent of n and b and are such that $E[\mathbf{X}] = \frac{\Theta(n^k)C^*}{n^l}$ and the equation is satisfied.

Then

$$\frac{A_F}{E[\mathbf{X}]^2} = \frac{E[\mathbf{X}]O\left(\frac{n^{k-f}}{n^{(2l-f\varepsilon'(H))b}}\right)}{E[\mathbf{X}]^2} = \frac{O\left(\frac{n^{k-f}}{n^{(2l-f\varepsilon'(H))b}}\right)}{E[\mathbf{X}]}$$

$$= \frac{O\left(\frac{n^{k-f}}{n^{(2l-f\varepsilon'(H))b}}\right)}{\Theta(n^{k-2l})} = O\left(n^{(\varepsilon'(H)b-1)f}\right).$$

Now, there are only a fixed number of subgraphs F of H, let this number be $s_H + 1$, so that there are s_H subgraphs of H that are not the empty graph F_0. So we have that

$$P[\mathbf{X} = 0] \leq \frac{\sum_F A_F - E[\mathbf{X}]^2}{E[\mathbf{X}]^2} = \frac{A_{F_0} + \sum_{F:F\neq F_0} A_F - E[\mathbf{X}]^2}{E[\mathbf{X}]^2}$$

$$\leq \frac{E[\mathbf{X}]^2 + \sum_{F:F\neq F_0} A_F - E[\mathbf{X}]^2}{E[\mathbf{X}]^2}$$

(since $A_{F_0} \leq E[\mathbf{X}]^2$)

$$= \sum_{F:F\neq F_0} \frac{A_F}{E[\mathbf{X}]^2} = \sum_{F:F\neq F_0} O\left(n^{(\varepsilon'(H)b-1)f}\right)$$

$$= s_H O\left(n^{(\varepsilon'(H)b-1)f}\right) = O\left(n^{(\varepsilon'(H)b-1)f}\right).$$

Since $b < \frac{1}{\varepsilon' H}$ and $f > 0$, we have that $(\varepsilon'(H)b - 1)f < 0$ and so $P[X = 0] \leq O\left(n^{(\varepsilon'(H)b-1)f}\right) \to 0$ as $n \to \infty$ whenever $b < \frac{1}{\varepsilon' H}$.

Therefore $b = \frac{1}{\varepsilon' H}$ is a threshold for the appearance of H.

QED.

3.2.2 Degree Distribution

We would like to show that the Sparse Random Dot Product Graph obeys a degree power law. However, similarly to the case of Dense Random Dot Product Graphs, the direct calculation of the degree distribution is difficult. So we again look at a slight variation of the idea and instead calculate the number of vertices that fall within a set interval.

Similar to the discussion in Chapter 2, for a given integer k and $\delta \in (0, 1)$ we wish to know the number of vertices in a graph with degree in the interval $[k(1-\delta), k(1+\delta)]$. We will notate this value as $\lambda[k(1 - \delta), k(1 + \delta)]$. Here, as in Chapter 2, we do not count the number of vertices whose degrees fall in $[k(1 - \delta), k(1 + \delta)]$ directly. Instead we select a value $s \in [0, 1]$ so that if $x_v = s$ then the $E[d(v)|x_v = s] =$

$\frac{n^{1-b}x_v}{a+1} = \frac{n^{1-b}s}{a+1} = k$. Then we count the number of vertices whose vectors fall in the interval $\mathcal{S} = [s(1-\delta), s(1+\delta)]$ since for any vertex v with $x_v \in \mathcal{S}$ then $E[d(v)|x_v] = \frac{n^{1-b}x_v}{a+1} \in \left[\frac{n^{1-b}s(1-\delta)}{a+1}, \frac{n^{1-b}s(1+\delta)}{a+1}\right] = [k(1-\delta), k(1+\delta)]$. And likewise, if $x_v \notin \mathcal{S}$ then $E[d(v)|x_v] \notin [k(1-\delta), k(1+\delta)]$. So, the number of vertices with $x_v \in \mathcal{S}$ is expected to be the same as the number of vertices with $d(v) \in [k(1-\delta), k(1+\delta)]$.

Now to allow for the variance that occurs in our model, we look at intervals $\mathcal{S}_-$ and $\mathcal{S}_+$ which are slightly smaller and larger than $\mathcal{S}$, respectively. We show that $\{v : x_v \in \mathcal{S}_-\} \subseteq \{v : d(v) \in [k(1-\delta), k(1+\delta)]\} \subseteq \{v : x_v \in \mathcal{S}_+\}$, each containment occurring with high probability. And therefore, the number of vertices whose degrees fall in $[k(1-\delta), k(1+\delta)]$ must be bounded by the number of vertices whose vectors fall in $\mathcal{S}_-$ and $\mathcal{S}_+$.

Finally, as in Chapter 2, we define an **X**-labeled Sparse Random Dot Product Graph. As with an ordinary Sparse Random Dot Product Graph, for each $v \in V(G)$ let the vector x_v be a one-dimensional vector drawn from $\mathcal{U}^a[0,1]$ and let the probability mapping $f = \frac{r}{n^b}$ where b is a positive real number. Additionally, let $\mathbf{X}$ be the $1 \times n$ matrix of vectors. We denote this sample space of **X**-labeled Sparse Random Dot Product Graph on n vertices as $\mathcal{DS}[\mathbf{X}, b]$. Note that the only difference between $G \in \mathcal{DS}[n, a, b, 1]$ and $(G, \mathbf{X}) \in \mathcal{DS}[\mathbf{X}, b]$ is that when G is a **X**-labeled Sparse Random Dot Product Graph the vectors are retained.

Lemma 3.2.2. *Let $(G, \mathbf{X})$ be drawn from $\mathcal{DS}[\mathbf{X}, b]$. Let $s \in (0,1)$,$\delta, \varepsilon_1 > 0$ and* *small. Define the interval* $\mathcal{S}_- = [\frac{s(1-\delta)}{1-\varepsilon_1}, \frac{s(1+\delta)}{1+\varepsilon_1}]$. *Then*

$$\mu_{\mathcal{S}_-} = E[d(v)|x_v \in \mathcal{S}_-] \in \left[\frac{n^{1-b}s(1-\delta)}{(1-\varepsilon_1)(a+1)}, \frac{n^{1-b}s(1+\delta)}{(1+\varepsilon_1)(a+1)}\right].$$

Proof:

$$\mu_{\mathcal{S}_-} = E[d(v)|x_v \in \mathcal{S}_-] = E\left[\sum_{w \in V(G)\,:\,w \neq v} \mathbf{I}\{v \sim w | x_v \in \mathcal{S}_-\}\right]$$

$$= (n-1)P[v \sim w | x_v \in \mathcal{S}_-]$$

$$= (n-1)\frac{\int_{\frac{s(1-\delta)}{1-\varepsilon_1}}^{\frac{s(1+\delta)}{1+\varepsilon_1}} \int_0^1 \frac{x_v x_u}{n^b} g(x_v) g(x_u) dx_v dx_u}{\int_{\frac{s(1-\delta)}{1-\varepsilon_1}}^{\frac{s(1+\delta)}{1+\varepsilon_1}} g(x_v) dx_v}$$

$$= \frac{(n-1)}{n^b} \frac{\left(\int_{\frac{s(1-\delta)}{1-\varepsilon_1}}^{\frac{s(1+\delta)}{1+\varepsilon_1}} x_v g(x_v) dx_v\right)\left(\int_0^1 x_u g(x_u) dx_u\right)}{\int_{\frac{s(1-\delta)}{1-\varepsilon_1}}^{\frac{s(1+\delta)}{1+\varepsilon_1}} g(x_v) dx_v}.$$

We look at each of the integrations separately. First note that $\int_0^1 x_u g(x_u) dx_u = \frac{1}{a+1}$. Next we note that

$$\int_{\frac{s(1-\delta)}{1-\varepsilon_1}}^{\frac{s(1+\delta)}{1+\varepsilon_1}} g(x_v) dx_v = \left[\left(\frac{s(1+\delta)}{1+\varepsilon_1}\right)^{\frac{1}{a}} - \left(\frac{s(1-\delta)}{1-\varepsilon_1}\right)^{\frac{1}{a}}\right] = \left[\hat{\delta}\frac{1}{a}s^{*\frac{1}{a}-1}\right]$$

for $\hat{\delta} = \left(\frac{s(1+\delta)}{1+\varepsilon_1}\right) - \left(\frac{s(1-\delta)}{1-\varepsilon_1}\right)$ and some $s^* \in [\left(\frac{s(1-\delta)}{1-\varepsilon_1}\right), \left(\frac{s(1+\delta)}{1+\varepsilon_1}\right)]$ by the mean value theorem.

Similarly

$$\int_{\frac{s(1-\delta)}{1-\varepsilon_1}}^{\frac{s(1+\delta)}{1+\varepsilon_1}} x_v g(x_v) dx_v = \frac{1}{1+a}\left[\hat{\delta}\frac{a+1}{a}s^{**\frac{1}{a}}\right]$$

for some $s^{**} \in [\left(\frac{s(1-\delta)}{1-\varepsilon_1}\right), \left(\frac{s(1+\delta)}{1+\varepsilon_1}\right)]$.

So we have

$$\mu_{\mathcal{S}_-} = \frac{(n-1)}{n^b(a+1)} \frac{\frac{1}{a+1}\left[\hat{\delta}\frac{a+1}{a}s^{**\frac{1}{a}}\right]}{\left[\hat{\delta}\frac{1}{a}s^{*\frac{1}{a}-1}\right]} = \frac{(n-1)}{n^b(a+1)} s^{*1-\frac{1}{a}} s^{**\frac{1}{a}} \sim \frac{n^{1-b}}{(a+1)} s^{*1-\frac{1}{a}} s^{**\frac{1}{a}}$$

since $\frac{n-1}{n^b} \sim n^{1-b}$.

Now $\left(\frac{s(1-\delta)}{1-\varepsilon_1}\right) \leq s^*, s^{**} \leq \left(\frac{s(1+\delta)}{1+\varepsilon_1}\right)$. Therefore $\left(\frac{s(1-\delta)}{1-\varepsilon_1}\right) \leq s^{*1-\frac{1}{a}} s^{**\frac{1}{a}} \leq \left(\frac{s(1+\delta)}{1+\varepsilon_1}\right)$.

So, we have

$$\frac{s(1-\delta)}{1-\varepsilon_1}\frac{n^b}{a+1} \leq \mu_{\mathcal{S}_-} \leq \frac{s(1+\delta)}{1+\varepsilon_1}\frac{n^b}{a+1}.$$

QED.

We have bounded the expected degree of any vertex whose vector falls in $\mathcal{S}_-$.

Lemma 3.2.3. *Let $(G, \mathbf{X})$ be drawn from $\mathcal{DS}[\mathbf{X}, b]$. Let $s, \mathcal{S}_-, \mu_{\mathcal{S}_-}, \delta$ and ε_1 be defined as in Lemma 3.2.2 with the additional condition that $(\varepsilon_1)^2\, s\, n^{1-b} \to \infty$. Let $v \in V(G)$. If $\frac{s(1-\delta)}{1-\varepsilon_1} \leq x_v \leq \frac{s(1+\delta)}{1+\varepsilon_1}$, then with probability tending to 1 as $n \to \infty$*

$$d(v) \in \left[\frac{s(1-\delta)n^{1-b}}{a+1}, \frac{s(1+\delta)n^{1-b}}{a+1}\right].$$

Indeed the probability $d(v) \notin \left[\frac{s(1-\delta)n^{1-b}}{a+1}, \frac{s(1+\delta)n^{1-b}}{a+1}\right]$ goes to zero faster than a reciprocal of any polynomial in n.

Proof:

Let $v \in V(G)$ and $\frac{s(1-\delta)}{1-\varepsilon_1} \leq x_v \leq \frac{s(1+\delta)}{1+\varepsilon_1}$. Now, the degree of a vertex $d(v)$ is the sum of the iid indicator variables $\mathbf{I}\{v \sim w\}$ and so by Chernoff's bounds we have $P[d(v) < (1-\varepsilon_1)\mu_{\mathcal{S}_-}] \leq \exp\left\{\frac{-\varepsilon_1^2 \mu_{\mathcal{S}_-}}{3}\right\}$. And so by Lemma 3.2.2 we have

$$P\left[d(v) < (1-\delta)s\frac{n^{1-b}}{a+1}\right] = P\left[d(v) < (1-\varepsilon_1)\frac{(1-\delta)}{1-\varepsilon_1}s\frac{n^{1-b}}{a+1}\right]$$

$$\leq P[d(v) < (1-\varepsilon_1)\mu_{\mathcal{S}_-}]$$

$$\leq \exp\left\{\frac{-\varepsilon_1^2 \mu_{\mathcal{S}_-}}{3}\right\}$$

$$\leq \exp\left\{\frac{-\varepsilon_1^2(1-\delta)(n^{1-b})s}{(1-\varepsilon_1)(a+1)3}\right\} \to 0$$

as $n \to \infty$ since $(\varepsilon_1)^2\, s\, n^{1-b} \to \infty$.

Similarly $P\left[d(v) > (1+\delta)s\frac{n-1}{a+1}\right] \leq \exp\left\{\frac{-\varepsilon_1^2(1-\delta)(n-1)s}{(1-\varepsilon_1)(a+1)3}\right\} \to 0$ as $n \to \infty$. And so with probability tending to 1 as $n \to \infty$

$$d(v) \in \left[\frac{s(1-\delta)n^{1-b}}{a+1}, \frac{s(1+\delta)n^{1-b}}{a+1}\right].$$

QED.

And so we have shown that with high probability if $x_v \in \mathcal{S}_-$ then $d(v) \in [k(1-\delta), k(1+\delta)]$ since $k = \frac{n^{1-b}s}{a+1}$.

Lemma 3.2.4. *Let $(G, \mathbf{X})$ be drawn from $\mathcal{DS}[\mathbf{X}, b]$. Let $s, \mathcal{S}_-, \mu_{\mathcal{S}_-}, \delta$ and ε_1 be defined* *as in Lemma 3.2.3 and* $k = \frac{s(a+1)}{n^{1-b}}$. *Let* $l > 0$, $\varepsilon_2 > 0$ *with the additional conditions that* $(1-\varepsilon_2)(1+\varepsilon_1) > 1$, $s(1-\delta) \geq l(1-\varepsilon_1)$, $(\varepsilon_1)^2\, n^{1-b}\, l \to \infty$ *and* $(\varepsilon_2)^2\, n^{1-b}\, s \to \infty$ *as* $n \to \infty$. *Define the interval* $\mathcal{S}_+ = [(1-\delta)s(1-\varepsilon_1), (1+\delta)s(1+\varepsilon_1)]$. *Let* $v \in V(G)$. *If* $k(1-\delta) \leq d(v) \leq k(1+\delta)$ *then with probability tending to* 1 *as* $n \to \infty$

$$x_v \in \mathcal{S}_+ = [(1-\delta)s(1-\varepsilon_1), (1+\delta)s(1+\varepsilon_1)].$$

Indeed the probability $x_v \notin \mathcal{S}_+$ *goes to zero faster than a reciprocal of any polynomial in* n.

Proof:

Let $d(v) \in [k(1-\delta), k(1+\delta)]$ First we examine what occurs when the vectors fall below the interval $\mathcal{S}_+$. By way of contradiction, assume $x_v < (1-\varepsilon_1)s(1-\delta)$.

Let's consider the case when $l \leq x_v < (1-\varepsilon_1)s(1-\delta)$. Recalling that $E[d(v)|x_v] = \frac{n^{1-b}x_v}{a+1}$, the Chernoff bound gives us that

$$P[d(v) > (1+\varepsilon_1)E[d(v)|x_v]] \leq \exp\left\{\frac{-\varepsilon_1^2 E[d(v)|x_v]}{3}\right\}$$

$$= \exp\left\{\frac{-\varepsilon_1^2 n^{1-b} x_v}{3(a+1)}\right\}$$

$$\leq \exp\left\{\frac{-\varepsilon_1^2 n^{1-b} l}{3(a+1)}\right\} \to 0$$

as $n \to \infty$ since $x_v \geq l$ and $(\varepsilon_1)^2\, n^{1-b}\, l \to \infty$.

Also, $x_v < (1-\varepsilon_1)s(1-\delta)$ gives us that

$$P[d(v) > (1+\varepsilon_2)E[d(v)|x_v]] = P\left[d(v) > (1+\varepsilon_1)\frac{n^{1-b}x_v}{a+1}\right]$$

$$\geq P[d(v) > (1+\varepsilon_1)\frac{n^{1-b}}{(a+1)}(1-\varepsilon_1)s(1-\delta)]$$

$$= P[d(v) > k(1+\varepsilon_1)(1-\varepsilon_1)(1-\delta)]$$

$$\geq P[d(v) \geq k(1-\delta)]$$

since $\varepsilon_1 > 0$ implies that $(1+\varepsilon_1)(1-\varepsilon_1) < 1$.

Therefore

$$P[d(v) \geq k(1-\delta)] \leq \exp\left\{\frac{-\varepsilon_1^2 n^{1-b} l}{3(a+1)}\right\}.$$

So,

$$E[\,|\{v : d(v) \geq k(1-\delta)\}||l \leq x_v < (1-\varepsilon_1)s(1-\delta)]$$

$$\leq n\,P[d(v) \geq k(1-\delta)\}||l \leq x_v < (1-\varepsilon_1)s(1-\delta)]$$

$$\leq n\exp\left\{\frac{-\varepsilon_1^2 n^{1-b} l}{3(a+1)}\right\} \to 0$$

as $n \to \infty$, since $\log n - \frac{\varepsilon_1^2 n^{1-b} l}{3(a+1)} \to -\infty$. Therefore by Markov's inequality

$$P[\,|\{v : d(v) \geq k(1-\delta)\}| = 0 \;\; |l \leq x_v < (1-\varepsilon_1)s(1-\delta)] \to 1.$$

So, if $l \leq x_v < (1-\varepsilon_1)s(1-\delta)$ then, with probability tending to 1, $d(v) < k(1-\delta)$

which is a contradiction.

Now consider the case when $0 \leq x_v \leq l$. In this case

$$E[d(v)|0 \leq x_v \leq l] = E\left[\sum_{w \in V(G)\,:\,w \neq v} \mathbf{I}\left\{v \sim w|0 \leq x_v \leq l\right\}\right]$$

$$= (n-1)P[v \sim w|0 \leq x_v \leq l]$$

$$= (n-1)\frac{\int_0^l \int_0^1 \frac{x_v x_w}{n^b} f(x_v)f(x_w)dx_v dx_w}{\int_0^l f(x_v)dx_v}$$

$$= \frac{(n-1)l}{n^b(a+1)^2}.$$

So the above equation and the fact that $s(1-\delta) \geq l(1-\varepsilon_1)$ give us that

$$P[d(v) > k(1-\delta)] \leq P\left[d(v) > \frac{n^{1-b}\,s}{(a+1)}(1-\delta)\right]$$

$$= P\left[d(v) > \frac{n^{1-b}\,s}{(a+1)^2}(1-\delta)\right]$$

(since $a > 0$)

$$\leq P\left[d(v) > \frac{n^{1-b}}{(a+1)^2} l(1-\varepsilon_1)\right]$$

$$= P[d(v) > (1+\varepsilon_1)E[d(v)|0 \leq x_v \leq n^{-1/4}]].$$

We see from the Chernoff bound that

$$P[d(v) > k(1-\delta)] \leq \exp\left\{\frac{-\varepsilon_1^2\,E[d(v)|0 \leq x_v \leq l]}{3}\right\}$$

$$= \exp\left\{\frac{-\varepsilon_1\, n^{1-b}\, l}{3(a+1)^2}\right\}.$$

So,

$$E[\,|\{v : d(v) \geq k(1-\delta)\}||0 \leq x_v \leq l]$$

$$\leq n\, P[d(v) \geq k(1-\delta)\}||0 \leq x_v \leq l]$$

$$\leq n \exp\left\{\frac{-\varepsilon_1\, n^{1-b}\, l}{3(a+1)^2}\right\} \to 0$$

as $n \to \infty$, since $\log n - (\frac{\varepsilon_1\, n^{1-b}\, l}{3(a+1)^2}) \to -\infty$. Therefore by Markov's inequality $P[\,|\{v : d(v) \geq k(1-\delta)\}| = 0] \to 1$. So, if $0 \leq x_v \leq l$ then, with probability tending to 1, $d(v) < k(1-\delta)$ which is again a contradiction.

Therefore, since both cases create a contradiction, with high probability $x_v \geq (1-\varepsilon_1)s(1-\delta)$ and the vectors will not fall below the interval $\mathcal{S}_+$.

Now we look for a contradiction when the vectors are above the interval $\mathcal{S}_+$. By way of contradiction, assume $x_v > (1+\varepsilon_1)s(1+\delta)$. The Chernoff bound gives

$$P[d(v) < (1-\varepsilon_2)E[d(v)|x_v]] \leq \exp\left\{\frac{-\varepsilon_2^2 E[d(v)|x_v]}{3}\right\}$$

$$= \exp\left\{\frac{-\varepsilon_2^2\, n^{1-b} x_v}{(a+1)3}\right\}$$

$$\leq \exp\left\{\frac{-\varepsilon_2^2\, n^{1-b}}{(a+1)3}(1+\varepsilon_1)s(1+\delta)\right\}$$

(since $x_v > (1+\varepsilon_1)s(1+\delta)$)

$$\leq \exp\left\{\frac{-\varepsilon_2^2\, n^{1-b} s}{(a+1)3}\right\}$$

(since $(1+\varepsilon_1)(1+\delta) > 1$).

Also,

$$P[d(v) < (1-\varepsilon_2)E[d(v)|x_v]] = P\left[d(v) > (1-\varepsilon_2)\frac{n^{1-b}x_v}{(a+1)}\right]$$

$$\geq P\left[d(v) < (1-\varepsilon_2)\frac{n^{1-b}}{(a+1)}(1+\varepsilon_1)s(1+\delta)\right]$$

(since $x_v > (1+\varepsilon_1)s(1+\delta)$)

$$= P[d(v) < k(1-\varepsilon_2)(1+\varepsilon_1)(1+\delta)]$$

$$\geq P[d(v) \leq k(1+\delta)]$$

since $(1-\varepsilon_2)(1+\varepsilon_1) > 1$. Finally we see that

$$P[d(v) \leq k(1+\delta)] \leq \exp\left\{\frac{-\varepsilon_2^2\, n^{1-b} s}{(a+1)3}\right\}.$$

So,

$$E[\,|\{v : d(v) \leq k(1+\delta)\}|\,|x_v > (1+\varepsilon_1)s(1+\delta)]$$

$$\leq n\, P[d(v) \leq k(1+\delta)\}||x_v > (1+\varepsilon_1)s(1+\delta)]$$

$$\leq \exp\left\{\frac{-\varepsilon_2^2\, n^{1-b} s}{(a+1)3}\right\} \to 0$$

as $n \to \infty$, since $\log n - (\frac{\varepsilon_2^2\, n^{1-b}\, s}{2(a+1)}) \to -\infty$ fast. Therefore by Markov's inequality $P[\,|\{v : d(v) \leq k(1+\delta)\}| = 0] \to 1$. So, if $x_v > (1+\varepsilon_1)s(1+\delta)$ then, with probability tending to 1, $d(v) > k(1+\delta)$ which is a contradiction. Therefore, probability tending to 1, $x_v \leq (1+\varepsilon_1)s(1+\delta)$ as $n \to \infty$.

QED.

And so we see that whenever a vertex has $d(v) \in [k(1-\delta), \leq\, k(1+\delta)]$ then $x_v \in \mathcal{S}_+$ with probability tending to 1 as $n \to \infty$.

Theorem 3.2.5. *Let* $(G, \mathbf{X})$ *be drawn from* $\mathcal{DS}[\mathbf{X}, b]$*. Let* $s \in (0,1)$*,* $k = \frac{s(a+1)}{n^{1-b}}$*,and* $\delta, \varepsilon_1, \varepsilon_2, \varepsilon_3, l > 0$ *with the following conditions*

- $(\varepsilon_1)^2\, n^{1-b}\, s \to \infty$
- $(\varepsilon_1)^2\, n^{1-b}\, l \to \infty$
- $(\varepsilon_2)^2\, n^{1-b}\, s \to \infty$
- $(\varepsilon_3)^2\, n\, \delta\, s^{\frac{1}{a}} \to \infty$
- $\varepsilon_1 \ll \delta$
- $s(1-\delta) \geq l(1-\varepsilon_1)$
- $(1-\varepsilon_2)(1+\varepsilon_1) > 1$

as $n \to \infty$.

Define $\mathcal{S}_- = [\frac{s(1-\delta)}{1-\varepsilon_1}, \frac{s(1+\delta)}{1+\varepsilon_1}]$ *and* $\mathcal{S}_+ = [(1-\varepsilon_1)s(1-\delta), (1+\varepsilon_1)s(1+\delta)]$. *Then* $\{v : x_v \in \mathcal{S}_-\} \subseteq \{v : d(v) \in [k(1-\delta), k(1+\delta)]\} \subseteq \{v : x_v \in \mathcal{S}_+\}$ *and*

$$\lambda[k(1-\delta), k(1+\delta)] = \left(\frac{n^{1-\frac{1}{a}-\frac{b}{a}}(a+1)^{\frac{1}{a}}}{a}\right)(2\,k[(\delta+\mathcal{O}(\varepsilon))(1+\mathcal{O}(\delta))^{\frac{1}{a}-1})]k^{\frac{1}{a}-1}(1+\mathcal{O}(\varepsilon_3))$$

with probability $\to 1$ *as* $n \to \infty$.

Proof:

Take $v \in \{v : x_v \in \mathcal{S}_-\}$. Note that all of the conditions of Lemma 3.2.3 are satisfied in the statement of the theorem. Therefore with probability tending to 1 as $n \to \infty$ we have that $d(v) \in \left[\frac{s(1-\delta)n^{1-b}}{a+1}, \frac{s(1+\delta)n^{1-b}}{a+1}\right] = [k(1-\delta), k(1+\delta)]$. Therefore $v \in \{v : d(v) \in [k(1-\delta), k(1+\delta)]\}$ and

$$P\left[\{v : x_v \in \mathcal{S}_-\} \subseteq \{v : d(v) \in [k(1-\delta), k(1+\delta)]\}\right] \to 1$$

as $n \to \infty$.

Similarly, take $v \in \{v : d(v) \in [k(1-\delta), k(1+\delta)]\}$. Also in the statement of the theorem, all of the conditions of Lemma 3.2.4 are satisfied. Therefore with probability tending to 1 as $n \to \infty$ we have that $x_v \in \mathcal{S}_+$ and so $v \in \{v : x_v \in \mathcal{S}_+\}$. Hence

$$P\left[\{v : d(v) \in [k(1-\delta), k(1+\delta)]\} \subseteq \{v : x_v \in \mathcal{S}_+\}\right] \to 1$$

as $n \to \infty$.

We have shown that with high probability $\{v : x_v \in \mathcal{S}_-\} \subseteq \{v : d(v) \in [k(1-\delta), k(1+\delta)]\} \subseteq \{v : x_v \in \mathcal{S}_+\}$ and so $|\{v : x_v \in \mathcal{S}_-\}| \leq \lambda[k(1-\delta), k(1+\delta)] \leq |\{v : x_v \in \mathcal{S}_+\}|$.

Now

$$E[|\{v : x_v \in \mathcal{S}_-\}|] = E\left[\sum_{v \in V(G)} \mathbf{I}\{x_v \in \mathcal{S}_-\}\right]$$

$$= nP[x_v \in \mathcal{S}_-]$$

$$= n \int_{\frac{s(1-\delta)}{1-\varepsilon_1}}^{\frac{s(1+\delta)}{1+\varepsilon_1}} g(x_v)dx_v$$

which we know from the proof of Lemma 3.2.2 to be

$$= n\left[\hat{\delta}\frac{1}{a}s^{*\frac{1}{a}-1}\right]$$

for $\hat{\delta} = \left(\frac{s(1+\delta)}{1+\varepsilon_1}\right) - \left(\frac{s(1-\delta)}{1-\varepsilon_1}\right)$ and some $s^* \in [\left(\frac{s(1-\delta)}{1-\varepsilon_1}\right), \left(\frac{s(1+\delta)}{1+\varepsilon_1}\right)]$.

Now, $\frac{1}{1-\varepsilon_1} = 1 + \varepsilon_1 + \frac{\varepsilon_1^2}{1-\varepsilon_1}$. So,

$$\frac{s(1-\delta)}{1-\varepsilon_1} = s(1-\delta)(1+\varepsilon_1+\frac{\varepsilon_1^2}{1-\varepsilon_1}) = s - s(\delta - \varepsilon_1 + \delta\varepsilon_1 + \frac{\delta\varepsilon_1^2}{1-\varepsilon_1} - \frac{\varepsilon_1^2}{1-\varepsilon_1})$$

$$= s - s\delta - s\,\mathcal{O}(\varepsilon_1).$$

Likewise $\frac{1}{1+\varepsilon_1} = 1 - \varepsilon_1 + \frac{\varepsilon_1^2}{1+\varepsilon_1}$ and

$$\frac{s(1+\delta)}{1+\varepsilon_1} = s + s\delta - s\,\mathcal{O}(\varepsilon_1).$$

Hence, we know that $\hat{\delta} = 2s\delta - s\,\mathcal{O}(\varepsilon_1)$. Also since $s, s^* \in [\left(\frac{s(1-\delta)}{1-\varepsilon_1}\right), \left(\frac{s(1+\delta)}{1+\varepsilon_1}\right)]$, we have that $s^* = s + s\,\mathcal{O}(\delta)$. So we can replace s^* giving us that

$$E[|\{v : x_v \in \mathcal{S}_-\}|] = n\left[\hat{\delta}\frac{1}{a}s^{*\frac{1}{a}-1}\right] = n\left[(2s\delta - s\,\mathcal{O}(\varepsilon_1))\frac{1}{a}(s + s\mathcal{O}(\delta))^{\frac{1}{a}-1}\right]$$

$$= n\left[(2\delta - \mathcal{O}(\varepsilon_1))\frac{1}{a}s^{\frac{1}{a}}(1+\mathcal{O}(\delta))^{\frac{1}{a}-1}\right]$$

and substituting $s = \frac{k(a-1)}{n^{1-b}}$ we have

$$= n\left[(2\delta - \mathcal{O}(\varepsilon_1))\frac{1}{a}\left(\frac{k(a-1)}{n^{1-b}}\right)^{\frac{1}{a}}(1+\mathcal{O}(\delta))^{\frac{1}{a}-1}\right]$$

$$= \left(\frac{n^{1-\frac{1}{a}-\frac{b}{a}}(a+1)^{\frac{1}{a}}}{a}\right)(2\,k[(\delta - \mathcal{O}(\varepsilon_1))(1+\mathcal{O}(\delta))^{\frac{1}{a}-1})]k^{\frac{1}{a}-1}.$$

For any $0 < \varepsilon_3 < 1$ the Chernoff bounds give us that

$$P[|\{v : x_v \in \mathcal{S}_-\}| < (1-\varepsilon_3)E[|\{v : x_v \in \mathcal{S}_-\}|] \leq \exp\left\{\frac{-\varepsilon_3^2 E[|\{v : x_v \in \mathcal{S}_-\}|]}{3}\right\}$$

$$= \exp\left\{\frac{-\varepsilon_3^2\left(\frac{n^{1-\frac{1}{a}-\frac{b}{a}}(a+1)^{\frac{1}{a}}}{a}\right)(2\,k[(\delta - \mathcal{O}(\varepsilon))(1+\mathcal{O}(\delta))^{\frac{1}{a}-1})]k^{\frac{1}{a}-1}}{3}\right\}$$

$$= \exp\left\{\frac{-\varepsilon_3^2 n\left[(2\delta - \mathcal{O}(\varepsilon_1))\frac{1}{a}s^{\frac{1}{a}}(1+\mathcal{O}(\delta))^{\frac{1}{a}-1}\right]}{3}\right\}$$

since $k = \frac{s(a+1)}{n^{1-b}}$.

We would like $P[|\{v : x_v \in \mathcal{S}_-\}| < (1 - n^{-1/3})E[|\{v : x_v \in \mathcal{S}_-\}|] \to 0$ as $n \to \infty$ and so we need

$$-\varepsilon_3^2 n\left[(2\delta - \mathcal{O}(\varepsilon_1))\frac{1}{a}s^{\frac{1}{a}}(1+\mathcal{O}(\delta))^{\frac{1}{a}-1}\right] \to \infty.$$

Now $\varepsilon_1 \ll \delta$ and so $2\delta - \mathcal{O}(\varepsilon_1) \sim 2\delta$. Also $(1+\mathcal{O}(\delta))^{\frac{1}{a}-1} \to 1$ as $n \to \infty$. Therefore we have that

$$P[|\{v : x_v \in \mathcal{S}_-\}| < (1-\varepsilon_3)E[|\{v : x_v \in \mathcal{S}_-\}|] \sim \exp\left\{\frac{-\varepsilon_3^2 n\left[2\delta\frac{1}{a}s^{\frac{1}{a}}\right]}{3}\right\} \to 0$$

as $n \to \infty$, since $\varepsilon_3^2\, n\, \delta\, s^{\frac{1}{a}} \to \infty$. Similarly

$$E[|\{v : x_v \in \mathcal{S}_+\}|] = \left(\frac{n^{1-\frac{1}{a}-\frac{b}{a}}(a+1)^{\frac{1}{a}}}{a}\right) 2\,k[(\delta + \mathcal{O}(\varepsilon))(1+\mathcal{O}(\delta))^{\frac{1}{a}-1})]k^{\frac{1}{a}-1}$$

and likewise

$$P[|\{v : x_v \in \mathcal{S}_+\}| > (1+\varepsilon_3)E[|\{v : x_v \in \mathcal{S}_+\}|] \to 0$$

as $n \to \infty$.

So we have that

$$(1-\varepsilon_3)\left(\frac{n^{1-\frac{1}{a}-\frac{b}{a}}(a+1)^{\frac{1}{a}}}{a}\right)(2\,k[(\delta-\mathcal{O}(\varepsilon))(1+\mathcal{O}(\delta))^{\frac{1}{a}-1})]k^{\frac{1}{a}-1} \leq \lambda[k(1-\delta),k(1+\delta)]$$

$$\leq (1+\varepsilon_3)\left(\frac{n^{1-\frac{1}{a}-\frac{b}{a}}(a+1)^{\frac{1}{a}}}{a}\right)(2\,k[(\delta+\mathcal{O}(\varepsilon))(1+\mathcal{O}(\delta))^{\frac{1}{a}-1})]k^{\frac{1}{a}-1}$$

with probability tending to 1 as $n \to \infty$. Therefore

$$\lambda[k(1-\delta),k(1+\delta)] = \left(\frac{n^{1-\frac{1}{a}-\frac{b}{a}}(a+1)^{\frac{1}{a}}}{a}\right)(2\,k[(\delta+\mathcal{O}(\varepsilon))(1+\mathcal{O}(\delta))^{\frac{1}{a}-1})]k^{\frac{1}{a}-1}(1+\mathcal{O}(\varepsilon_3))$$

with probability $\to 1$ as $n \to \infty$.

QED.

So, as long as we can satisfy all of the conditions of Theorem 3.2.5 we are able to show that $\lambda[k(1-\delta),k(1+\delta)]$ is proportional to $(2k\delta)k^{\frac{1-a}{a}}$ with high probability. And we are closer to showing that the degree distribution is indeed power law. The question becomes "Can we satisfy all of the conditions of the theorem?"

Theorem 3.2.6. *Whenever $0 < b < 1$ the conditions of Theorem 3.2.5 are satisfiable.*

Proof:

Let $0 < b < 1$. Let $s \in (0,1)$, $k = \frac{s(a+1)}{n^{1-b}}$,and $\delta, \varepsilon_1, \varepsilon_2, \varepsilon_3, l > 0$. We would like to satisfy the following conditions:

1. $(\varepsilon_1)^2\, n^{1-b}\, s \to \infty$

2. $(\varepsilon_1)^2\, n^{1-b}\, l \to \infty$

3. $(\varepsilon_2)^2\, n^{1-b}\, s \to \infty$

4. $(\varepsilon_3)^2\, n\, \delta\, s^{\frac{1}{a}} \to \infty$

5. $\varepsilon_1 \ll \delta$

6. $s(1-\delta) \geq l(1-\varepsilon_1)$

7. $(1-\varepsilon_2)(1+\varepsilon_1) > 1$

as $n \to \infty$.

Let us consider when $\delta = n^{-x_\delta}$, $l = n^{-x_l}$, $s = n^{-x_s}$ and for $i = 1, 2, 3$, $\varepsilon_i = n^{-x_i}$. Then we can rewrite the conditions as:

1. $n^{-2x_1}\, n^{1-b}\, n^{x_s} = n^{1-b-(2x_1+x_s)} \to \infty$

2. $n^{-2x_1}\, n^{1-b}\, n^{x_l} = n^{1-b-(2x_1+x_l)} \to \infty$

3. $n^{-2x_2}\, n^{1-b}\, n^{x_s} = n^{1-b-(2x_2+x_s)} \to \infty$

4. $n^{-2x_3}\, n\, n^{-x_\delta}\, (n^{-x_s})^{\frac{1}{a}} = n^{1-(2x_3+x_\delta+\frac{x_s}{a})} \to \infty$

5. $n^{x_1} \ll n^{-x_\delta}$

6. $n^{-x_s}(1-n^{-x_\delta}) \geq n^{-x_l}(1-n^{-x_1})$

7. $(1-n^{-x_2})(1+n^{-x_1}) > 1$

as $n \to \infty$. We can again rewrite the first five conditions as:

1. $2x_1 + x_s < 1 - b$

2. $2x_1 + x_l < 1 - b$

3. $2x_2 + x_s < 1 - b$

4. $2x_3 + x_\delta + \frac{x_s}{a} < 1$

5. $x_1 > x_\delta$

Now let us consider condition 6. Since $\varepsilon_1 > 0$ we know that $1 - \varepsilon_1 < 1$ therefore if we require that $n^{-x_s}(1 - n^{-x_\delta}) \geq n^{-x_l}$ we will still satisfy condition 6. For large n, $n^{-x_\delta} < \frac{1}{2}$ and $n^{-x_s}(1 - n^{-x_\delta}) \geq \frac{1}{2}n^{-x_s}$, so if we require that $\frac{1}{2}n^{-x_s} > n^{-x_l}$ condition 6 will still be satisfied. For large n, $\frac{1}{2}n^{-x_s} > n^{-x_l}$ is satisfied whenever $x_l > x_s$ and this is our new condition 6.

6. $x_l > x_s$

Finally, let us consider condition 7, $(1 - n^{-x_2})(1 + n^{-x_1}) > 1$ is equivalent to $n^{-x_1} > n^{-x_2}(1 + n^{-x_1})$. Also $1 + n^{-x_1} < 2$, therefore we will satisfy condition 7 as long as $n^{-x_1} > 2n^{-x_2}$ which for large n will be satisfied as long as $x_2 > x_1$. So our new condition 7 is

7. $x_2 > x_1$

and our conditions become a simple system of linear inequalities.

If $a > 1$ let us fix a value $D = \frac{1-b}{8}$ and let $x_1 = 2D$, $x_2 = 3D$, $x_3 = D$, $x_l = 2D$, $x_s = D$, and $x_\delta = D$. Then we see that all 7 conditions of the theorem are satisfied.

If $a < 1$ let $D = a\frac{1-b}{8}$ and the rest be as above and again all 7 conditions of the Theorem 3.2.5 are satisfied.

QED.

Hence, the conditions of Theorem 3.2.5 are satisfiable and it will work for values such as $k = \frac{n^{a-b}s}{a+1} \geq \frac{n^{\frac{5}{6}(1-b)}}{a+1}$.

3.2.3 Diameter

Our goal is to show that the Sparse Random Dot Product Graph is connected (barring isolated vertices) and determine the diameter of this connected component. To this end we apply the following approach. We establish a core of vertices with a known small diameter. Then we select a threshold value so that whenever a vector is greater than the threshold, the corresponding vertex is almost surely adjacent to a vertex in the core. Finally, we show that whenever we have two adjacent vertices whose vectors fall below the threshold, then at least one of them must be adjacent to a vertex in the core. Here, as in Section 3.2.2, we will consider the $\mathbf{X}$-labeled Sparse Random Dot Product Graph.

Lemma 3.2.7. *Let $(G, \mathbf{X})$ be drawn from $\mathcal{DS}[\mathbf{X}, b]$ with $b \in (0, 1)$. Let $\varepsilon > 0$ and let $d > \frac{1}{1-2\varepsilon-b}$. Let $\mathcal{C}$ be the subgraph of G induced by $V(\mathcal{C}) = \{v \in V(G) : x_v \geq n^{-\varepsilon}\}$. Then $P[diam(\mathcal{C}) > d] \to 0$ as $n \to \infty$.*

Proof:

Let $u, v \in V(\mathcal{C})$, then $P[u \sim v] \geq \frac{n^{-2\varepsilon}}{n^b} = \frac{1}{n^{2\varepsilon+b}}$. In [14], Klee and Larman showed that for a fixed d^*, the Erdős-Rényi random graph with edge probability p has diameter at most d^* with probability tending to 1 if $\frac{(pn)^{d^*}}{n} \to \infty$. Therefore in our model we know that the $diam(\mathcal{C}) < d$ (with high probability) whenever $\frac{\left((\frac{1}{n^{2\varepsilon+b}})n\right)^d}{n} \to \infty$. However we are given that $d > \frac{1}{1-2\varepsilon-b}$ and so we see that $(1-2\varepsilon-b)d - 1 > (1-2\varepsilon-b)\frac{1}{1-2\varepsilon-b} - 1 = 0$. Hence

$$\frac{\left((\frac{1}{n^{2\varepsilon+b}})n\right)^d}{n} = n^{(1-2\varepsilon-b)d-1} \to \infty.$$

QED.

So $\mathcal{C}$ is a low diameter core of the Sparse Random Dot Product Graph.

Given a graph G, subgraph H of G, and vertex $v \in V(G) - V(H)$, we say that $v \sim H$ if for any $u \in V(H)$, $v \sim u$ and $v \not\sim H$ if for all $u \in V(H)$, $v \not\sim u$.

Lemma 3.2.8. *Let $(G, \mathbf{X})$ be drawn from $\mathcal{DS}[\mathbf{X}, b]$ with $b \in (0,1)$ and $\mathcal{C}$ and ε be as defined in Lemma 3.2.7. Let $0 < g < 1 - b - \varepsilon$. Let $\hat{V} = \{v : x_v \geq n^{-g}\}$, then $\forall v \in \hat{V}$ $P[v \not\sim \mathcal{C}] \to 0$ as $n \to \infty$. Furthermore, $P[|\{v \in \hat{V}$ with $v \not\sim \mathcal{C}| = 0] \to 1$ as $n \to \infty$.*

In other words, whenever a vertex v has vector $x_v \geq n^{-g}$ then v has an edge to some vertex in the core $\mathcal{C}$.

Proof:

Let $v \in \hat{V}$. Then

$$P[v \nsim \mathcal{C}] \leq \prod_{u \in \mathcal{C}} P[v \nsim u] = \prod_{u \in \mathcal{C}} (1 - P[v \sim u])$$

$$\leq \prod_{u \in \mathcal{C}} \left(1 - \frac{n^{-g} n^{-\varepsilon}}{n^b}\right) = \left(1 - \frac{1}{n^{b+g+\varepsilon}}\right)^{|V(\mathcal{C})|}.$$

Let us recall that for any vertex $w \in V(G)$, $P[x_w \geq n^{-\varepsilon}] = (1-(n^{-\varepsilon})^{\frac{1}{a}})$. Therefore $E[|V(\mathcal{C})|] = n(1 - n^{\frac{-\varepsilon}{a}})$. Also, by Chernoff's inequality, the number of such vertices is, with high probability, at least $0.9n(1 - n^{\frac{-\varepsilon}{a}})$ and so we see that

$$P[v \nsim \mathcal{C}] \leq \left(1 - \frac{1}{n^{b+g+\varepsilon}}\right)^{0.9n(1-n^{\frac{-\varepsilon}{a}})}.$$

Now, $b + g + \varepsilon < 1$, therefore $\exists\, \delta > 0$ such that $b + g + \varepsilon + \delta = 1$ and

$$P[v \nsim \mathcal{C}] \leq \left(1 - \frac{n^{\delta}}{n}\right)^{0.9n(1-n^{\frac{-\varepsilon}{a}})} \sim e^{-n^{\delta}} \to 0$$

as $n \to \infty$, since $\delta > 0$.

Furthermore, $E[|\{v \in \hat{V} \text{ with } v \nsim \mathcal{C}|] \leq nP[v \nsim \mathcal{C} | v \in \hat{V}] \sim ne^{-n^{\delta}} \to 0$ as $n \to \infty$. Therefore by Markov's inequality, $P[|\{v \in \hat{V} \text{ with } v \nsim \mathcal{C}| = 0] \to 1$ as $n \to \infty$.

QED.

Lemma 3.2.9. *Let* $(G, \mathbf{X})$ *be drawn from* $\mathcal{DS}[\mathbf{X}, b]$ *and* $\mathcal{C}, \varepsilon$ *and* g *be as defined in Lemma 3.2.8 with the additional condition that* $\varepsilon \ll \frac{\frac{2}{a+2} - b}{1 + \frac{2}{a}}$ *and* $0 < b < \frac{2}{a+2}$. *Let* $S = \{\{u, v\} : x_u, x_v < n^{-g}, u \sim v,\ u \not\sim \mathcal{C}, v \not\sim \mathcal{C}\}$, *then* $P[|S| = 0] \to 1$ *as* $n \to \infty$.

Proof:

Let u, v be such that $x_u, x_v < n^{-g}$. Then

$$P[u \sim v,\ u \not\sim \mathcal{C}, v \not\sim \mathcal{C} | x_u, x_v] = \frac{x_u x_v}{n^b} \prod_{y \in V(\mathcal{C})} \left(1 - \frac{x_u x_y}{n^b}\right)\left(1 - \frac{x_v x_y}{n^b}\right)$$

$$\leq \frac{x_u x_v}{n^b} \prod_{y \in V(\mathcal{C})} \left(1 - \frac{x_u n^{-\varepsilon}}{n^b}\right)\left(1 - \frac{x_v n^{-\varepsilon}}{n^b}\right)$$

$$= \frac{x_u x_v}{n^b} \left(1 - \frac{x_u n^{-\varepsilon}}{n^b}\right)^{|V(\mathcal{C})|} \left(1 - \frac{x_v n^{-\varepsilon}}{n^b}\right)^{|V(\mathcal{C})|}$$

and recalling from Lemma 3.2.8 that $|V(\mathcal{C})| \geq 0.9n(1 - n^{\frac{-\varepsilon}{a}})$ we have that

$$\leq \frac{x_u x_v}{n^b} \left(1 - \frac{x_u n^{-\varepsilon}}{n^b}\right)^{0.9n(1-n^{\frac{-\varepsilon}{a}})} \left(1 - \frac{x_v n^{-\varepsilon}}{n^b}\right)^{0.9n(1-n^{\frac{-\varepsilon}{a}})}.$$

Now we integrate to remove the conditioning on x_u, x_v and we have

$$P[u \sim v,\ u \not\sim \mathcal{C}, v \not\sim \mathcal{C} | x_u, x_v < n^{-g}] = \frac{P[u \sim v,\ u \not\sim \mathcal{C}, v \not\sim \mathcal{C}, x_u, x_v < n^{-g}]}{P[x_u, x_v < n^{-g}]}$$

$$= \frac{1}{n^{\frac{-2g}{a}}} \int_0^{n^{-g}} \int_0^{n^{-g}} \frac{x_u x_v}{n^b} \left(1 - \frac{x_u n^{-\varepsilon}}{n^b}\right)^{0.9n(1-n^{\frac{-\varepsilon}{a}})}$$

$$\cdot\left(1-\frac{x_v n^{-\varepsilon}}{n^b}\right)^{0.9n(1-n^{\frac{-\varepsilon}{a}})}\frac{1}{a}x_u{}^{\frac{1-a}{a}}dx_u\,\frac{1}{a}x_v{}^{\frac{1-a}{a}}dx_v$$

$$=\frac{1}{n^{\frac{-2g}{a}+b}}\left(\int_0^{n^{-g}}x\left(1-\frac{x}{n^{b+\varepsilon}}\right)^{0.9n(1-n^{\frac{-\varepsilon}{a}})}\frac{1}{a}x^{\frac{1-a}{a}}dx\right)^2.$$

Let $t=x^{\frac{1}{a}}$, so that $dt=\frac{1}{a}x^{\frac{1-a}{a}}dx$ for $t\in(0,n^{-\frac{g}{a}})$ and substituting we have

$$P[u\sim v,\ u\not\sim\mathcal{C},v\not\sim\mathcal{C}|x_u,x_v<n^{-g}]=\frac{1}{n^{\frac{-2g}{a}+b}}\left(\int_0^{n^{-\frac{g}{a}}}t^a\left(1-\frac{t^a}{n^{b+\varepsilon}}\right)^{0.9n(1-n^{\frac{-\varepsilon}{a}})}dt\right)^2$$

$$\leq\frac{1}{n^{\frac{-2g}{a}+b}}\left(\int_0^1 t^a\left(1-\frac{t^a}{n^{b+\varepsilon}}\right)^{0.9n(1-n^{\frac{-\varepsilon}{a}})}dt\right)^2$$

since $n^{-\frac{g}{a}}<1$ for large n and the integrand is positive.

Now let $t=\frac{y}{n^{\frac{1-(b+\varepsilon)}{a}}}$, so that $dt=\frac{dy}{n^{\frac{1-(b+\varepsilon)}{a}}}$ for $y\in(0,n^{\frac{1-(b+\varepsilon)}{a}})$ and again substituting we have

$$P[u\sim v,\ u\not\sim\mathcal{C},v\not\sim\mathcal{C}|x_u,x_v<n^{-g}]\leq$$

$$\leq\frac{1}{n^{\frac{-2g}{a}+b}}\left(\int_0^{n^{\frac{1-(b+\varepsilon)}{a}}}\frac{y^a}{n^{1-b-\varepsilon}}\left(1-\frac{y^a}{n}\right)^{0.9n(1-n^{\frac{-\varepsilon}{a}})}\frac{dy}{n^{\frac{1-(b+\varepsilon)}{a}}}\right)^2$$

$$=\frac{1}{n^{\frac{-2g}{a}+b+2(1-b-\varepsilon)+2(\frac{1-b-\varepsilon}{a})}}\left(\int_0^{n^{\frac{1-(b+\varepsilon)}{a}}}y^a\left(1-\frac{y^a}{n}\right)^{0.9n(1-n^{\frac{-\varepsilon}{a}})}dy\right)^2$$

and from Lemma 3.2.8 we know that $b+g+\varepsilon<1$ implying that $\frac{1-(b+\varepsilon)}{a}>0$ and

$$\sim\frac{1}{n^{\frac{-2g}{a}+b+2(1-b-\varepsilon)+2(\frac{1-b-\varepsilon}{a})}}\left(\int_0^{\infty}y^a\,e^{-0.9y^a}dy\right)^2$$

$$= \frac{1}{n^{\frac{-2g}{a}+b+2(1-b-\varepsilon)+2(\frac{1-b-\varepsilon}{a})}} \left(\frac{\Gamma(\frac{1}{a})}{(0.9)^{\frac{a+1}{a}} a^2} \right)^2$$

$$= \frac{\Theta(1)}{n^{\frac{-2g}{a}+b+2(1-b-\varepsilon)+2(\frac{1-b-\varepsilon}{a})}}.$$

Now, for any vertex $v \in V(G)$, $P[x_v < n^{-g}] = n^{-\frac{g}{a}}$. Therefore

$$E[|\{v \in V(G) : x_v < n^{-g}\}|] = n\, n^{-\frac{g}{a}}.$$

So we have the following result

$$E[|S|] = \binom{E[|\{v \in V(G) : x_v < n^{-g}\}|]}{2} P[u \sim v,\ u \not\sim \mathcal{C},\ v \not\sim \mathcal{C} | x_u, x_v < n^{-g}]$$

$$\leq \frac{n^2 n^{-\frac{2g}{a}} \Theta(1)}{n^{\frac{-2g}{a}+b+2(1-b-\varepsilon)+2(\frac{1-b-\varepsilon}{a})}}$$

$$= \frac{\Theta(1)}{n^{-b-2\varepsilon+\frac{2}{a}-\frac{2b}{a}-\frac{2\varepsilon}{a}}} \to 0$$

as $n \to \infty$ since $b < \frac{2}{a+2}$ and $\varepsilon \ll \frac{\frac{2}{a+2}-b}{1+\frac{2}{a}}$. Therefore $E[|S|] \to 0$ as $n \to \infty$ and by Markov's inequality we have that $P[|S| = 0] \to 1$.

QED.

Hence, with high probability, if $\exists\ u, v \in V(G) - V(\mathcal{C})$ with $u \sim v$ then either $u \sim \mathcal{C}$ or $v \sim \mathcal{C}$.

Now, as a result of Lemmas 3.2.8 and 3.2.9, with high probability, any vertex that is not isolated is either adjacent to $\mathcal{C}$ or has a neighbor that is adjacent to $\mathcal{C}$. Therefore with high probability, the probability that G is made up of one large component $\hat{G}$ and isolated vertices tends to one as n goes to ∞. With this in mind we obtain the following result.

Theorem 3.2.10. *Let $(G, \mathbf{X})$ be drawn from $\mathcal{DS}[\mathbf{X}, b]$. Let $0 < \varepsilon \ll \frac{\frac{2}{a+2} - b}{1+\frac{2}{a}}$, $0 < g < 1 - b - \varepsilon$, and $0 < b < \frac{2}{a+2}$. Let $\hat{G}$ be the subgraph of G induced by $V(\hat{G}) = \{v \in V(G) : d(v) > 0\}$. Then for all $d > \frac{1}{1-2\varepsilon-b}$, $P[diam(\hat{G}) \leq d+4] \to 1$ as $n \to \infty$.*

Proof:

Choose any two vertices $u, v \in V(\hat{G})$. Without loss of generality assume that $x_u \leq x_v$. We have the following six cases.

Case 1: $x_u, x_v \geq n^{-\varepsilon}$

By Lemma 3.2.7, we have $P[d(u, v) \leq d] \to 1$ as $n \to \infty$.

Case 2: $n^{-g} \leq x_u < n^{-\varepsilon}, x_v \geq n^{-\varepsilon}$

By Lemma 3.2.8 with probability $\to$ one $u \sim \mathcal{C}$ and $\exists\, y \in \mathcal{C}$ such that $u \sim y$. Also, $d(u, v) \leq d(u, y) + d(y, v) = 1 + d(y, v)$ and so by Lemma 3.2.7 as $n \to \infty$

$$P[d(u, v) \leq d+1] \geq P[1 + d(y, v) \leq d+1] = P[d(y, v) \leq d] \to 1.$$

Case 3: $n^{-g} \leq x_u, x_v < n^{-\varepsilon}$

By Lemma 3.2.8 with probability $\to$ one, $u \sim \mathcal{C}$ and $v \sim \mathcal{C}$. Therefore $\exists\, y \in \mathcal{C}$ and $z \in \mathcal{C}$ such that $u \sim y$ and $v \sim z$. Also, $d(u, v) \leq d(u, y) + d(y, z) + d(z, v) = 2 + d(y, z)$.

Then by Lemma 3.2.7 we have that

$$P[d(u,v) \leq d+2] \geq P[2+d(y,z) \leq d+2] = P[d(y,z) \leq d] \to 1.$$

Case 4: $x_v < n^{-g}, x_v \geq n^{-\varepsilon}$

Now, since $\hat{G}$ is connected $\exists\, y \in \hat{G}$ such that $u \sim y$ and $d(u,v) \leq d(u,y)+d(y,v) = 1+d(y,v)$. If $x_y \geq n^{-\varepsilon}$ then by Lemma 3.2.7 as $n \to \infty$,

$$P[d(u,v) \leq d+1] \geq P[1+d(y,v) \leq d+1] = P[d(y,v) \leq d] \to 1.$$

If $n^{-g} \leq x_y < n^{-\varepsilon}$ then by Lemma 3.2.8, $y \sim \mathcal{C}$ and $\exists\, z \in \mathcal{C}$ such that $y \sim z$. Also, $d(u,v) \leq d(u,y)+d(y,z)+d(z,v) = 2+d(z,v)$. Therefore by Lemma 3.2.7 we have that

$$P[d(u,v) \leq d+2] \geq P[d(z,v)+2 \leq d+2] = P[d(z,v) \leq d] \to 1.$$

If $x_y < n^{-g}$, then by Lemma 3.2.9 with probability $\to$ one either u or y will be adjacent to a vertex z in $\mathcal{C}$. Hence $d(u,z) \leq 2$ and so $d(u,v) \leq d(u,z)+d(z,v) \leq 2+d(z,v)$. Therefore by Lemma 3.2.7 we have that

$$P[d(u,v) \leq d+2] \geq P[2+d(z,v) \leq d+2] \geq P[d(z,v) \leq d] \to 1$$

as $n \to \infty$.

Case 5: $x_u < n^{-g}, n^{-g} \leq x_v < x^{-\varepsilon}$

First, since $n^{-g} \leq x_v < x^{-\varepsilon}$ we know by Lemma 3.2.8 that $\exists\, w \in V(\mathcal{C})$ such that $v \sim w$.

Also, since $\hat{G}$ is connected $\exists\ y \in \hat{G}$ such that $u \sim y$ and $d(u,v) \leq d(u,y) + d(y,w) + d(w,v) = 2 + d(y,w)$. If $x_y \geq n^{-\varepsilon}$ then by Lemma 3.2.7 as $n \to \infty$,

$$P[d(u,v) \leq d+2] \geq P[2 + d(y,w) \leq d+2] = P[d(y,w) \leq d] \to 1.$$

If $n^{-g} \leq x_y < n^{-\varepsilon}$ then by Lemma 3.2.8, $y \sim \mathcal{C}$ and $\exists\, z \in V(\mathcal{C})$ such that $y \sim z$. Also, $d(u,v) \leq d(u,y) + d(y,z) + d(z,w) + d(w,v) = 3 + d(z,w)$. Therefore by Lemma 3.2.7 we have that

$$P[d(u,v) \leq d+3] \geq P[d(z,w) + 3 \leq d+3] = P[d(z,w) \leq d] \to 1.$$

If $x_y < n^{-g}$, then by Lemma 3.2.9 with probability $\to$ one either u or y will be adjacent to a vertex z in $\mathcal{C}$. Hence $d(u,z) \leq 2$ and so $d(u,v) \leq d(u,z) + d(z,w) + d(w,v) \leq 3 + d(z,w)$. Therefore by Lemma 3.2.7 we have that

$$P[d(u,v) \leq d+3] \geq P[3 + d(z,v) \leq d+3] \geq P[d(z,v) \leq d] \to 1$$

as $n \to \infty$.

Case 6: $x_u, x_v < n^{-g}$

In this case we are going to establish upper bounds on the distances from u and

v to the core $\mathcal{C}$. Then we will calculate an upper bound the on the distance from u to v.

First, since $\hat{G}$ is connected $\exists\ y \in \hat{G}$ such that $u \sim y$. Now, if $x_y < n^{-g}$, then by Lemma 3.2.9 with probability $\rightarrow$ one either u or y will be adjacent to a vertex $\hat{y}$ in $\mathcal{C}$. Hence $d(u, \hat{y}) \leq 2$. If $x_y \geq n^{-g}$ then by Lemma 3.2.8, $y \sim \mathcal{C}$ and $\exists\ \hat{y} \in V(\mathcal{C})$ such that $y \sim \hat{y}$. Here also $d(u, z) \leq d(u, y) + d(y, z) = 2$ and so in either case u is at most distance 2 from some vertex z in $\mathcal{C}$.

Since $\hat{G}$ is connected $\exists\ z \in \hat{G}$ such that $v \sim y$ and by discussion identical to that above there must also exist a $\hat{z}$ in $\mathcal{C}$ with distance at most 2 from v. So, $d(u, v) \leq d(u, \hat{y}) + d(\hat{y}, \hat{z}) + d(\hat{z}, v) \leq d(\hat{y}, \hat{z}) + 4$. Therefore by Lemma 3.2.7 we have that

$$P[d(u, v) \leq d + 4] \geq P[d(\hat{y}, \hat{z}) + 4 \leq d + 4] = P[d(\hat{y}, \hat{z}) \leq d] \rightarrow 1$$

as $n \rightarrow \infty$.

Therefore for all cases as $n \rightarrow \infty$, $P[d(u, v) \leq d + 4] \rightarrow 1$ and so $P[diam(\hat{G}) \leq d + 4] \rightarrow 1$.

QED.

3.3 The Complete Picture

In this section we summarize the results from the rest of this chapter.

First, consider the evolution of possible subgraphs as the parameter b goes from zero to infinity. Although we proved thresholds for individual subgraphs, the main theorem that aides us in this discussion is Theorem 3.2.1, which states that for any non-edgeless subgraph H, the threshold for the appearance of H is $b = \frac{1}{\epsilon'(H)}$. Therefore we know that $b = 2$ is the threshold for the appearance of edges and hence when $b > 2$ the Sparse Random Dot Product Graph G is almost surely edgeless. We also know that when $b = 0$ (the dense case) all subgraphs H are present with high probability.

Now let us consider subgraphs of order $k \in \mathbb{Z}_+$. We have the following, with high probability as we move along from $b = 0$ to infinity:

- beginning at $b = 0$ all subgraphs on k vertices are present,
- at $b = \frac{2}{k-1}$ cliques of size k disappear,
- at $b = 1$ k-cycles disappear,
- at $b = \frac{k}{k-1}$ trees on k vertices disappear,
- for $b > \frac{k}{k-1}$ there are no connected subgraphs on k vertices, and
- for $b > 2$ there are no edges in the graph.

Next let us review what we have shown with respect to the connectivity of the Sparse Random Dot Product Graph when $b \in (0, 1)$. We know that the Sparse Random Dot Product Graph always has isolated vertices (Theorem 3.1.10) and hence is never connected. However, if we ignore isolated vertices, Theorem 3.2.10 tells us that whenever $0 < b < \frac{\frac{2}{a}}{1+\frac{2}{a}}$, the diameter of the main connected component (again ignoring isolated vertices) is $d_b = \lfloor \frac{1}{1-2\epsilon-b} \rfloor + 5$.

Finally, we know a little about the degree distribution of the Sparse Random Dot Product Graph when $b \in (0,1)$. Proposition 3.1.8 states that for any fixed $k \in \mathbb{Z}_+$, $\lambda(k)$, the number of vertices of degree k has expected value $E[\lambda(k)] \sim C_{k,a} n^{\frac{a-1+b}{a}}$ for some constant $C_{k,a}$ not dependent on n. Additionally, by Theorem 3.2.5 we know that the degree distribution of the Sparse Random Dot Product Graph follows a power law.

So in summary, the Sparse Random Dot Product Graph has (with high probability) the following characteristics for the various values of b:

- When $b \in (0,1)$, all cycles and trees of finite size are present and the Sparse Random Dot Product Graph has clique number $\omega = \lceil \frac{2}{b} \rceil$ or $\lceil \frac{2}{b} \rceil + 1$ (by Corollary 3.1.6). Also, the graph always has isolated vertices, with the remaining vertices connected in a giant component of diameter $d_b = \lfloor \frac{1}{1-2\epsilon-b} \rfloor + 5$ whenever $b < \frac{\frac{2}{a}}{1+\frac{2}{a}}$. Finally, the graph obeys the degree power law.

- When $b \in (1,2)$, the Sparse Random Dot Product Graph is a forest with tree components of size $\frac{b}{b-1}$ or less and hence has $\omega = 2$.

- When $b \in (2,\infty)$, the Sparse Random Dot Product Graph is edgeless.

Chapter 4

Posynomial Asymptotics: A Prelude to the The Discrete Model

4.1 The β, α Parameter Space

In the next chapter we consider a discrete version of the Random Dot Product Graph. In this chapter, we develop a set of tools to facilitate the discussion of the discrete version.

Let $m, n, h > 0$ be integers. Assume that we have a polynomial in p and $\frac{1}{t}$ of the form

$$f(p, \frac{1}{t}) = n^h \left(a_0 p^{x_0} + a_1 \frac{p^{x_1}}{t^1} + a_2 \frac{p^{x_2}}{t^2} + \cdots + a_{m-1} \frac{p^{x_{m-1}}}{t^{m-1}} \right)$$

where $a_i \in \{0, 1\}$, but not all a_i are zero, and $0 < x_i \leq 2m$ for all $0 \leq i \leq m-1$.

Also, let us assume that $p = \frac{1}{n^\alpha}$ and $t = n^\beta$, where $\alpha, \beta \geq 0$, so that

$$f_{\alpha,\beta,h}(n) = \sum_{i=0}^{m-1} \frac{a_i(\frac{1}{n^\alpha})^{x_i}}{(n^\beta)^i} n^h = \sum_{i=0}^{m-1} a_i n^{h - x_i\alpha - \beta i}$$

and is a posynomial in n. We would like to know for which values of α and β we have $f_{\alpha,\beta,h}(n) \to 0$ or $f_{\alpha,\beta,h}(n) \to \infty$ as $n \to \infty$. Specifically we investigate the boundary between these two cases.

Now, $f_{\alpha,\beta,h}(n) \to 0$ as $n \to \infty$ if and only if every term in the sum goes to 0. Also, if any term in the sum goes to infinity, then $f_{\alpha,\beta,h}(n) \to \infty$ as $n \to \infty$, regardless of the behavior of the other terms. Consider the ith term of the sum. If $a_i = 0$ then the entire term is 0 and nothing is added to the summation. So, let us assume that $a_i = 1$ and the ith term of the sum is $n^{h-x_i\alpha-\beta i}$. Now, $n^{h-x_i\alpha-\beta i} \to 0$ if and only if $h - x_i\alpha - \beta i < 0$, and $n^{h-x_i\alpha-\beta i} \to \infty$ if and only if $h - x_i\alpha - \beta i > 0$. So, $h - x_i\alpha - \beta i = 0$ is the threshold for the ith term.

Define the β, α-parameter space to be the first quadrant of the Cartesian plane with α values on the vertical axis and β values on the horizontal axis. Then the threshold $h - x_i\alpha - \beta i = 0$ is the line, l_i that divides the parameter space into two pieces and we know the following:

- Fact 1: $l_i : h - x_i\alpha - \beta i = 0$ is a line with negative slope $\frac{-i}{x_i}$, α-intercept $\frac{h}{x_i}$ and β-intercept $\frac{h}{i}$.

- Fact 2: The origin $(0, 0)$ is not on the line $l_i : h - x_i\alpha - \beta i = 0$ since $h > 0$

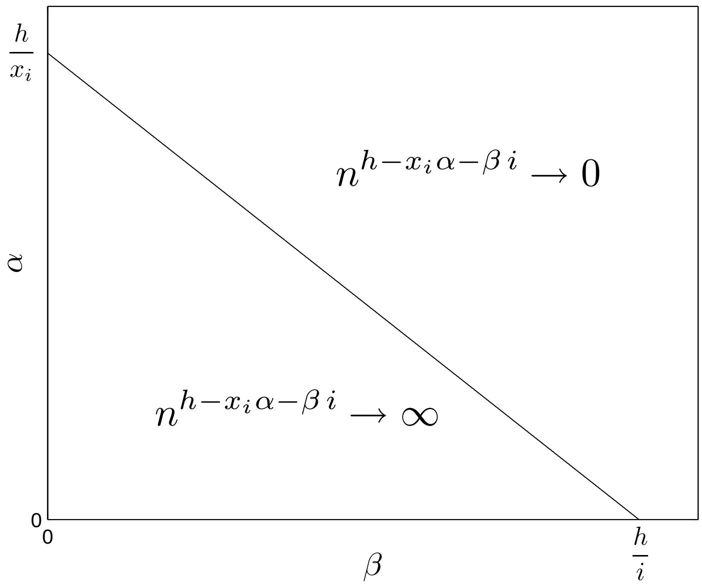


Figure 4.1: The line $l_i : h - x_i\alpha - \beta\, i = 0$.

and hence for all points (β, α) in the open half space containing the origin $h - x_i\alpha - \beta i > 0$ and the term $n^{h-x_i\alpha-\beta i} \to \infty$ as $n \to \infty$. We call this open half space containing the origin the open half space 'below' the line l_i.

- Fact 3: Similarly, for all points (β, α) in the open half space not containing the origin $h - x_i\alpha - \beta i < 0$ and the term $n^{h-x_i\alpha-\beta i} \to 0$ as $n \to \infty$. We call this open half space not containing the origin the open half space 'above' the line l_i.

For the remainder of this chapter, whenever we discuss half spaces and their intersections we are only referring to the part of the region that is within the β, α-parameter space, i.e., the part of the region for which $\alpha, \beta \geq 0$.

We study the behavior of $f_{\alpha,\beta,h}(n)$ with respect to this β, α-parameter space and so to this end we have the following proposition.

Proposition 4.1.1. *Let $m, h > 0$ be integers. Let $f_{\alpha,\beta,h}(n) = \sum_{i=0}^{m-1} a_i n^{h-x_i\alpha-\beta i}$ where $a_i \in \{0,1\}$, but not all zero, and $0 < x_i \leq 2m$ for all $0 \leq i \leq m-1$. Each term with $a_i \neq 0$ determines an open half space in the β, α-parameter space that does not contain the origin. The intersection of these 'above' open half spaces creates a convex set, $\mathcal{C}(f)$, whose points all yield $f_{\alpha,\beta,h}(n) \to 0$ as $n \to \infty$. Additionally, for all points not in this convex set or on its boundary, $f_{\alpha,\beta,h}(n) \to \infty$ as $n \to \infty$. We will call these points, not in $\mathcal{C}(f)$ or on its boundary, the corner of f and denote the set as $\zeta(f)$.*

Proof:

Each $0 \leq i \leq m-1$ for which $a_i \neq 0$ is associated with a line l_i in the parameter space in which the ith term goes to infinity in the open half space 'below' the line and to 0 in the open half space 'above' the line. Consider the intersection of all of the open half spaces 'above' these lines, $\mathcal{C}(f)$. For all points in $\mathcal{C}(f)$ the ith term of the summation, $0 \leq i \leq m-1$, has either $a_i = 0$ and is equal to 0, or has $a_i = 1$ and the ith term goes to 0 as $n \to \infty$ since the points are 'above' the line l_i. Therefore, for all points (β, α) in $\mathcal{C}(f)$, $f_{\alpha,\beta,h}(n) \to 0$ as $n \to \infty$.

For any point (β, α) in the corner of f, $\zeta(f)$, there is at least one term for which $a_i = 1$ and the point falls in the half space 'below' the line l_i, otherwise it would be contained in the convex set $\mathcal{C}(f)$. Therefore, that term goes to infinity as $n \to \infty$ and so $f_{\alpha,\beta,h}(n) \to \infty$ as well.

QED.

We call the piecewise linear function that defines the convex set $\mathcal{C}(f)$ the boundary of the of the set and refer to it as the function $\alpha = F(\beta)$.

For example, let $f_{\alpha,\beta,h}(n) = n^{5-14\alpha-3\beta} + n^{5-9\alpha-4\beta} + n^{5-6\alpha-5\beta} + n^{5-5\alpha-6\beta}$. Then Figure 4.2 (a) illustrates all of the lines associated with the terms of $f_{\alpha,\beta,h}(n)$ and Figure 4.2 (b) shows the convex region $\mathcal{C}(f)$ and the corner $\zeta(f)$ described in Proposition 4.1.1.

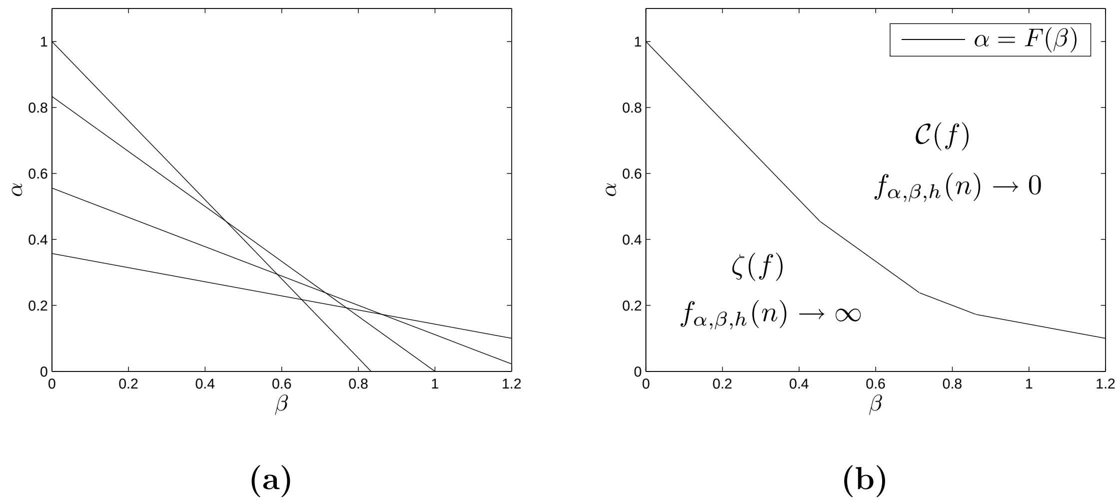


Figure 4.2: **(a)** The lines associated with $f_{\alpha,\beta,h}(n)$. **(b)** $\mathcal{C}(f)$ and $\zeta(f)$.

Definition 4.1.2. *Let $(\beta_1, \alpha_1), (\beta_2, \alpha_2)$ be points in the β, α-parameter space. We define a partial order as follows: $(\beta_1, \alpha_1) \le (\beta_2, \alpha_2)$ if and only if $\beta_1 \le \beta_2$ and $\alpha_1 \le \alpha_2$.*

Proposition 4.1.3. *Let $m, h > 0$ be integers. Let $f_{\alpha,\beta,h}(n) = \sum_{i=0}^{m-1} a_i n^{h-x_i\alpha-\beta i}$ where $a_i \in \{0, 1\}$, not all zero, and $0 < x_i \le 2m$ for all $0 \le i \le m-1$. Suppose that for the point (β_0, α_0) we know that $f_{\alpha_0,\beta_0,h}(n) \to 0$ as $n \to \infty$. Then for all points*

$(\beta, \alpha) \geq (\beta_0, \alpha_0)$, we have that $f_{\alpha,\beta,h}(n) \to 0$ as $n \to \infty$.

Proof:

Let (β_0, α_0) be such that $f_{\alpha_0,\beta_0,h}(n) \to 0$ as $n \to \infty$. Then in each term of the summation either $a_i = 0$ or the term goes to 0 as $n \to \infty$. Hence, whenever $a_i \neq 0$ we have that $h - x_i\alpha_0 - \beta_0 i < 0$. Now consider a point $(\beta, \alpha) \geq (\beta_0, \alpha_0)$. Then

$$h - x_i(\alpha) - i(\beta) \leq h - x_i\alpha_0 - \beta_0 i < 0$$

since $x_i, i \geq 0$, $\alpha \geq \alpha_0$, and $\beta \geq \beta_0$. Therefore, in each term of the summation evaluated at (β, α), either $a_i = 0$ or the term goes to 0 as $n \to \infty$ and so, $f_{\alpha,\beta,h}(n) \to 0$ as $n \to \infty$.

QED.

4.2 The Equivalence Classes

Let $\mathcal{F}$ be the set of all functions of the form $f_{\alpha,\beta,h}(n) = \sum_{i=0}^{m-1} a_i n^{h-x_i\alpha-\beta i}$ where $m, h \in \mathbb{Z}_+$,$a_i \in \{0, 1\}$, but not all a_i are zero, and $0 < x_i \leq 2m$ for all $0 \leq i \leq m-1$. Then by Proposition 4.1.1, each $f_{\alpha,\beta,h}(n) \in \mathcal{F}$ determines a convex set $\mathcal{C}(f)$ whose boundary is a piecewise linear function, $\alpha = F(\beta)$, and for which all points 'above' the boundary $f_{\alpha,\beta,h}(n) \to 0$ and for all points 'below' the boundary $f_{\alpha,\beta,h}(n) \to \infty$. Let $[f_{\alpha,\beta,h}(n)]$ be the set of all $g_{\alpha,\beta,h}(n) \in \mathcal{F}$ whose associated convex hull is the same as that of $f_{\alpha,\beta,h}(n)$. Clearly, having-the-same-convex-hull-as is an equivalence relation

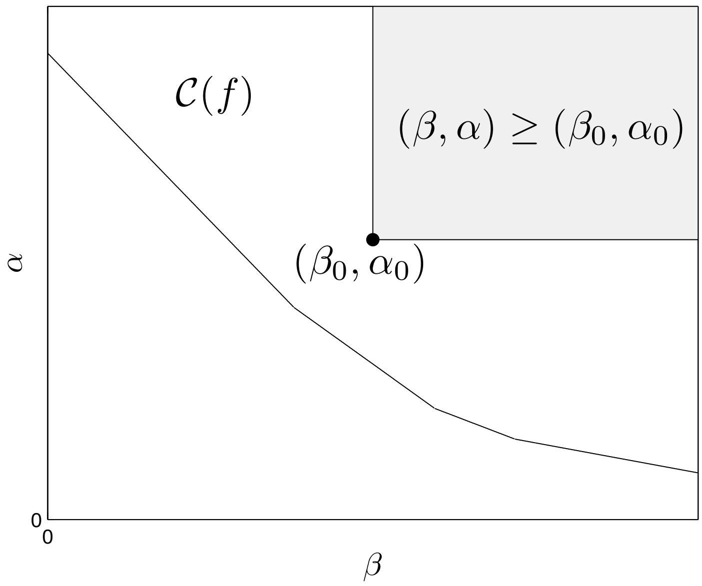


Figure 4.3: If $(\beta_0, \alpha_0) \in \mathcal{C}(f)$ then all points $(\beta, \alpha) \geq (\beta_0, \alpha_0)$ are in $\mathcal{C}(f)$.

on $\mathcal{F}$.

Proposition 4.2.1. *Let $\mathcal{F}$ be the set of all functions of the form*

$$f_{\alpha,\beta,h}(n) = \sum_{i=0}^{m-1} a_i n^{h - x_i\alpha - \beta i}$$

where $m, h \in \mathbb{Z}_+$,$a_i \in \{0, 1\}$, but not all a_i are zero, and $0 < x_i \leq 2m$ for all $0 \leq i \leq m-1$. In the context of Proposition 4.1.1, let $f_{\alpha,\beta,h}(n) \equiv g_{\alpha,\beta,h}(n) \in \mathcal{F}$ if the associated convex sets are the same, i.e., if $\mathcal{C}(f) = \mathcal{C}(g)$. Then, the property of 'having the same convex set', $\equiv$, is an equivalence relation on $\mathcal{F}$.

Proposition 4.2.2. *For all $f_{\alpha,\beta,h}(n), g_{\alpha,\beta,h}(n) \in \mathcal{F}$ the following are equivalent:*

- $f_{\alpha,\beta,h}(n) \equiv g_{\alpha,\beta,h}(n)$

- $g_{\alpha,\beta,h}(n) \in [f_{\alpha,\beta,h}(n)]$

- $G(\beta) = F(\beta)$

- $\mathcal{C}(g) = \mathcal{C}(f)$

- $g_{\alpha,\beta,h}(n) \to \infty$ *if and only if* $f_{\alpha,\beta,h}(n) \to \infty$ *and* $g_{\alpha,\beta,h}(n) \to 0$ *if and only if* $f_{\alpha,\beta,h}(n) \to 0$.

The proof Proposition of 4.2.2 is simple, based on definitions and Proposition 4.1.1, and is omitted here.

Consider a $f_{\alpha,\beta,h}(n) = \sum_{i=0}^{m-1} a_i n^{h-x_i\alpha-\beta i} \in \mathcal{F}$. Let $\mathcal{F}_{f_{\alpha,\beta,h}(n)}$ be the set of all $f'_{\alpha,\beta,h}(n) \in \mathcal{F}$ for which

$$f'_{\alpha,\beta,h}(n) = \sum_{i=0}^{m-1} a'_i n^{h-x_i\alpha-\beta i}$$

where $a'_i \leq a_i$. In other words, the ith term of $f'_{\alpha,\beta,h}(n)$ agrees with the ith term of $f_{\alpha,\beta,h}(n)$ or is 0. We would like to know for which $f'_{\alpha,\beta,h}(n) \in \mathcal{F}_{f_{\alpha,\beta,h}(n)}$, do we also have $f'_{\alpha,\beta,h}(n) \equiv f_{\alpha,\beta,h}(n)$.

Proposition 4.2.3. *Let $m, h > 0$ be integers. Let $f_{\alpha,\beta,h}(n) = \sum_{i=0}^{m-1} a_i n^{h-x_i\alpha-\beta i}$ where $a_i \in \{0,1\}$, not all zero, and $0 < x_i \leq 2m$ for all $0 \leq i \leq m-1$. Let $\alpha = F(\beta)$ be the boundary of the convex set $\mathcal{C}(f)$. If there exist $j \in \{0, m-1\}$ for which $a_j = 1$ and either*

- *the line l_j falls completely in the open region 'below' the convex intersection determined by $\alpha = F(\beta)$ or*

- *the line l_j falls in the open region 'below' the convex set intersecting $\alpha = F(\beta)$ at exactly one point*

then the intersection of the 'above' open half planes determined by the function $f'_{\alpha,\beta,h}(n) = \sum_{\substack{i=0 \\ i \neq j}}^{m-1} a_i n^{h - x_i\alpha - \beta i}$ is the same as that determined by $f_{\alpha,\beta,h}(n)$, i.e., $f'_{\alpha,\beta,h}(n) \equiv f_{\alpha,\beta,h}(n)$.

The proof is straightforward and is omitted.

Proposition 4.2.4. *Let $m, h > 0$ be integers. Let $f_{\alpha,\beta,h}(n) = \sum_{i=0}^{m-1} a_i n^{h - x_i\alpha - \beta i}$ where $a_i \in \{0, 1\}$, not all zero, and $0 < x_i \leq 2m$ for all $0 \leq i \leq m - 1$. Let $\alpha = F(\beta)$ be the boundary of the convex set $\mathcal{C}(f)$. Let $i_1 < i_2 < \cdots < i_k \in \{0, m-1\}$, $2 \leq k \leq m-1$, be such that $l_{i_1}, l_{i_2}, \ldots, l_{i_k}$ are all the same line. Then the intersection of the 'above' open half planes determined by the function*

$$f'_{\alpha,\beta,h}(n) = \sum_{\substack{i=0 \\ i \neq i_2,\ldots,i_k}}^{m-1} a_i n^{h - x_i\alpha - \beta i}$$

is the same as that determined by $f_{\alpha,\beta,h}(n)$, i.e., $f'_{\alpha,\beta,h}(n) \equiv f_{\alpha,\beta,h}(n)$.

Proof:

Let $i_1 < i_2 < \cdots < i_k \in \{0, m-1\}$, $2 \leq k \leq m-1$, be such that $l_{i_1}, l_{i_2}, \ldots, l_{i_k}$ are all the same line. Therefore, the 'above' and 'below' open half spaces determined by each of the $l_{i_1}, l_{i_2}, \ldots, l_{i_k}$ are the same. Hence the intersection of all k lines with the lines from the remaining terms is identical to the intersection of just l_{i_1} with the lines of the remaining terms. So, the function $f'_{\alpha,\beta,h}(n) = \sum_{\substack{i=0 \\ i \neq i_2,\ldots,i_k}}^{m-1} a_i n^{h - x_i\alpha - \beta i}$ is the same as the intersection of 'above' open half spaces determined $f_{\alpha,\beta,h}(n)$ and thus, $f'_{\alpha,\beta,h}(n) \equiv f_{\alpha,\beta,h}(n)$.

QED.

Propositions 4.2.3 and 4.2.4 give us insight into which $f'_{\alpha,\beta,h}(n) \in \mathcal{F}_{f_{\alpha,\beta,h}(n)}$, satisfy $f'_{\alpha,\beta,h}(n) \equiv f_{\alpha,\beta,h}(n)$. We take this idea further and ask for which $f'_{\alpha,\beta,h}(n) = \sum_{i=0}^{m-1} a'_i n^{h-x_i\alpha-\beta i} \in \mathcal{F}_{f_{\alpha,\beta,h}(n)}$ do we minimize the sum $\sum_{i=0}^{m-1} a'_i$ subject to equivalence with $f_{\alpha,\beta,h}(n)$.

Proposition 4.2.5. *Let $f_{\alpha,\beta,h}(n) = \sum_{i=0}^{m-1} a_i n^{h-x_i\alpha-\beta i} \in \mathcal{F}$ and $\alpha = F(\beta)$ be the boundary of the associated convex set as described in Proposition 4.1.1. Let $\mathcal{F}_{f_{\alpha,\beta,h}(n)}$ be the set of all $f'_{\alpha,\beta,h}(n) \in \mathcal{F}$ for which $f'_{\alpha,\beta,h}(n) = \sum_{i=0}^{m-1} a'_i n^{h-x_i\alpha-\beta i}$ where $a'_i \leq a_i$. Let*

$$f^{red}_{\alpha,\beta,h}(n) = \sum_{i=0}^{m-1} r'_i n^{h-x_i\alpha-\beta i}$$

be obtained from $f_{\alpha,\beta,h}(n)$ as follows:

- *Step 1: Let $i = 0$.*
- *Step 2: If $a_i = 0$, then set $r_i = 0$ and go to Step 5, otherwise if $a_i = 1$ go to Step 3.*
- *Step 3: If line l_i falls 'below' the boundary $\alpha = F(\beta)$ or intersects it in at most one point, set $r_i = 0$ and go to Step 5, otherwise go to Step 4.*
- *Step 4: If line l_i is the same as another line l_j where $j < i$, set $r_i = 0$ and go*

to Step 5, otherwise set $r_i = a_i = 1$ and go to Step 5.

- *Step 5: Set $i = i + 1$. If $i < m - 1$ go to Step 2, otherwise Stop.*

Then $f^{red}_{\alpha,\beta,h}(n) \in \mathcal{F}_{f_{\alpha,\beta,h}(n)}$ and $f^{red}_{\alpha,\beta,h}(n) \in [f_{\alpha,\beta,h}(n)]$. Also, for any $f'_{\alpha,\beta,h}(n) \in \mathcal{F} \bigcap [f_{\alpha,\beta,h}(n)]$, $\sum_{i=0}^{m-1} a_i' \geq \sum_{i=0}^{m-1} r_i$.

Proof:

First, for all $0 \leq i \leq m-1$, if $a_i = 0$ then $r_i = 0$. Also, if $a_i = 1$ then $r_i \in \{0, 1\}$. Therefore for all $0 \leq i \leq m-1$, $r_i \leq a_i$ and clearly we have that $f^{red}_{\alpha,\beta,h}(n) \in \mathcal{F}_{f_{\alpha,\beta,h}(n)}$.

Next we wish to show that $f^{red}_{\alpha,\beta,h}(n) \in [f_{\alpha,\beta,h}(n)]$. We do this by examining the steps of the reduction to see how they affect the convex set $\mathcal{C}(f^{red})$. First note that if for all $0 \leq i \leq m-1$, $r_i = a_i$ then clearly the two functions have the same convex set. So, we only need to be concerned with steps in the reduction that make $r_i \neq a_i$. In Step 3, we set $r_i = 0$ while $a_i = 1$. However this is done because the line l_i falls 'below' the boundary $\alpha = F(\beta)$ or intersects it in at most one point. Therefore by Proposition 4.2.3 the line l_i creates an 'above' open half plane that completely contains $\mathcal{C}(f)$ and removing the ith term from the summation does not change the convex set. Therefore, changing r_i to 0 does not affect associated convex set. Similarly, in Step 4 we set $r_i = 0$ while $a_i = 1$. However, this is done because the line l_i is identical to another line l_j with $j < i$. Therefore, by Proposition 4.2.4 we can remove this term without changing $\mathcal{C}(f)$. Therefore, changing r_i to 0 does not affect the associated convex set. Hence, the changes made to the r_i do not affect the convex set and so

$f^{red}_{\alpha,\beta,h}(n) \equiv f_{\alpha,\beta,h}(n)$. Therefore by Proposition 4.2.2, $f^{red}_{\alpha,\beta,h}(n) \in [f_{\alpha,\beta,h}(n)]$.

Finally we show that for any $f^{'}_{\alpha,\beta,h}(n) \in \mathcal{F} \bigcap [f_{\alpha,\beta,h}(n)]$, $\sum_{i=0}^{m-1} a_i^{'} \geq \sum_{i=0}^{m-1} r_i$. For the sake of contradiction, let us assume that there exists a function $f^{'}_{\alpha,\beta,h}(n) \in \mathcal{F} \bigcap [f_{\alpha,\beta,h}(n)]$ for which $\sum_{i=0}^{m-1} a_i^{'} = A < \sum_{i=0}^{m-1} r_i = R$. Hence A lines are used to determine the convex set associated with $f^{'}_{\alpha,\beta,h}(n)$, $\mathcal{C}(f^{'})$, and R lines are used to determine the convex set associated with $f^{red}_{\alpha,\beta,h}(n)$, $\mathcal{C}(f^{red})$. However, $\mathcal{C}(f^{'}) = \mathcal{C}(f^{red})$. Hence the piecewise linear functions associated with them are equal, i.e., $F^{'}(\beta) = F^{red}(\beta)$. Now, each linear piece of this function is a segment of one of the lines l_i associated with the terms of the function. So, if A lines are used to determine $F^{'}(\beta)$, only A lines should be needed to determine $F^{red}(\beta)$. This would imply that at least one of the R lines, say l_*, associated with $f^{red}_{\alpha,\beta,h}(n)$ is not needed to determine the convex set. This happens if l_* falls completely 'below' the boundary of the convex set, intersects it in at most one point, or falls on the boundary $F^{red}(\beta)$. If l_* falls completely 'below' the boundary of the convex set or intersect it at at most one point, then in Step 3 of the reduction, r_* would have been set to 0 and hence this line would not appear. If l_* falls on the boundary $F^{red}(\beta)$, but is not required to determine the convex set, then it must be equal to another line, say l_{**} and in Step 4 of the reduction either r_* or r_{**} would have been set to 0 and one of the lines would not be considered. Hence, all R lines are required to determine the piecewise linear function $F^{red}(\beta) = F^{'}(\beta)$ and so $A = R$ which is a contradiction. Therefore $\sum_{i=0}^{m-1} a_i^{'} \geq \sum_{i=0}^{m-1} r_i$.

QED.

Proposition 4.2.6. *Let $m, h > 0$ be integers. Let $f_{\alpha,\beta,h}(n) = \sum_{i=0}^{m-1} a_i n^{h-x_i\alpha-\beta i}$ where $a_i \in \{0,1\}$ and $0 < x_i \leq 2m$ for all $0 \leq i \leq m-1$. Let us have the additional assumption that $x_i > x_j$ whenever $i < j$. Let $\alpha = F(\beta)$ be the boundary of the convex set in which $f_{\alpha,\beta,h}(n) \to 0$ as described in Prop. 4.1.1. Consider the function $g_{\alpha,\beta,h}(n) = f_{\alpha,\beta,h}(n)^2$. Then the convex intersection of the 'above' open half planes determined by the function $g_{\alpha,\beta,h}(n)$ is the same as that determined by $f_{\alpha,\beta,h}(n)$, i.e., $g_{\alpha,\beta,h}(n) \equiv f_{\alpha,\beta,h}(n)$.*

Proof:

$$g_{\alpha,\beta,h}(n) = f_{\alpha,\beta,h}(n)^2 = \left(\sum_{i=0}^{m-1} a_i n^{h-x_i\alpha-\beta i}\right)^2$$

$$= \sum_{i=0}^{m-1} a_i^2 n^{2(h-x_i\alpha-\beta i)} + \sum_{\substack{i,j=0 \\ i\neq j}}^{m-1} a_i a_j n^{2h-\alpha(x_i+x_j)-\beta(i+j)}$$

As we discussed in the case of $f_{\alpha,\beta,h}(n)$, $g_{\alpha,\beta,h}(n) \to 0$ as $n \to \infty$ if and only if each term in the sum is or goes to 0. Also, if a single term of the sum goes to infinity then $g_{\alpha,\beta,h}(n) \to \infty$ as $n \to \infty$. So we examine the terms of $g_{\alpha,\beta,h}(n)$ individually.

First let us consider the terms of $g_{\alpha,\beta,h}(n)$ generated by the first summation. For any $0 \leq i \leq m-1$ we have that the ith term of the first summation is $a_i^2 n^{2(h-x_i\alpha-\beta i)}$. If $a_i = 0$ then the ith term of the summation is also 0, as is the ith term of $f_{\alpha,\beta,h}(n)$. So assume that $a_i = 1$, and the ith term of the first summation is $n^{2(h-x_i\alpha-\beta i)}$. The ith term of the first summation will go to 0 whenever $2(h - x_i\alpha - \beta i) < 0$ and will go to infinity whenever $2(h - x_i\alpha - \beta i) > 0$. So, the threshold for the ith term is the line $2(h - x_i\alpha - \beta i) = 0$. This line is exactly the same line as that defined by the ith term of $f_{\alpha,\beta,h}(n)$, $h - x_i\alpha - \beta i = 0$. Therefore, each of the terms in the first summation of $g_{\alpha,\beta,h}(n)$ determines open half spaces identical to the corresponding terms in $f_{\alpha,\beta,h}(n)$. Hence the intersection of the 'above' open half spaces is exactly that as defined by $\alpha = F(\beta)$ and any further intersection with 'above' open half spaces from the second summation will be contained in this set. Therefore without any further examination of terms, we know that the region defined by $\alpha = G(\beta)$ will be contained in that defined by $\alpha = F(\beta)$, i.e., $\mathcal{C}(g) \subseteq \mathcal{C}(f)$.

Now let us consider one of the cross terms in the second summation. Let $0 \leq i, j \leq m - 1$ with $i \neq j$ and consider the term $a_i a_j n^{2h-\alpha(x_i+x_j)-\beta(i+j)}$. If a_i or a_j is zero, the term is zero, so assume that $a_i = a_j = 1$. This i, jth term goes to 0 whenever $2h-\alpha(x_i+x_j)-\beta(i+j) < 0$ and goes to infinity whenever $2h-\alpha(x_i+x_j)-\beta(i+j) > 0$. So, the line

$$l_{i,j} : 2h - \alpha(x_i + x_j) - \beta(i + j) = 0$$

defines the threshold for this term. We show that the intersection of this open half space 'above' is the same as if the half space was not included and the i, jth term can be ignored. By Proposition 4.2.3 it is sufficient to show that the line $l_{i,j}$ falls entirely below $\alpha = F(\beta)$ or intersects the boundary function at at most one point.

Let us consider individually the ith and jth terms of the first summation. The ith and jth terms determine half spaces with the lines $l_i : 2h - 2x_i\alpha - 2\beta i = 0$ and

$l_j : 2h - 2x_j\alpha - 2\beta j = 0$, respectively, as their boundaries. Without loss of generality let us assume that $i < j$, and so $x_i > x_j$. Now, the line

$$l_i : 2h - 2x_i\alpha - 2\beta i = 0$$

intersects the α and β axes at the points $(0, \frac{h}{x_i})$ and $(\frac{h}{i}, 0)$, respectively, and has slope $\frac{-i}{x_i}$. Similarly, the line

$$l_j : 2h - 2x_j\alpha - 2\beta j = 0$$

intersects the α and β axes at the points $(0, \frac{h}{x_j})$ and $(\frac{h}{j}, 0)$, respectively, and has slope $\frac{-j}{x_j}$. Additionally, $x_i > x_j$ implies that $\frac{h}{x_i} < \frac{h}{x_j}$ and we see that the α-intercept of l_j is above that of l_i. Also, $i < j$ implies that $\frac{h}{i} > \frac{h}{j}$ and the β-intercept of l_i is to the right of the β-intercept of l_j. Therefore the two lines intersect within the β, α-parameter space and their intersection point is $(\frac{h(x_i - x_j)}{x_i j - x_j i}, \frac{h(j-i)}{x_i j - x_j i})$.

Let us return to the i, jth term. Recall that the line associated with the i, jth term is $l_{i,j} : 2h - \alpha(x_i + x_j) - \beta(i + j) = 0$. This line has α and β intercepts $(0, \frac{2h}{x_i + x_j})$ and $(\frac{2h}{i+j}, 0)$, respectively. These intercepts are the harmonic means of the α and β intercepts of the ith and jth terms and therefore fall between the intercepts on each axis, i.e. $\frac{h}{x_i} < \frac{2h}{x_i + x_j} < \frac{h}{x_j}$ and $\frac{h}{i} > \frac{2h}{i+j} > \frac{h}{j}$. Additionally, the point of intersection for the lines l_i and l_j, $(\frac{h(x_i - x_j)}{x_i j - x_j i}, \frac{h(j-i)}{x_i j - x_j i})$, falls on the line $l_{i,j}$. So, we see that, when $\beta \in [0, \frac{h(x_i - x_j)}{x_i j - x_j i})$, $l_{i,j}$ falls 'below' l_j and when $\beta \in (\frac{h(x_i - x_j)}{x_i j - x_j i}, \infty)$, $l_{i,j}$ falls 'below' l_i, with the three lines intersecting at a single point when $\beta = \frac{h(x_i - x_j)}{x_i j - x_j i}$. Hence the $l_{i,j}$

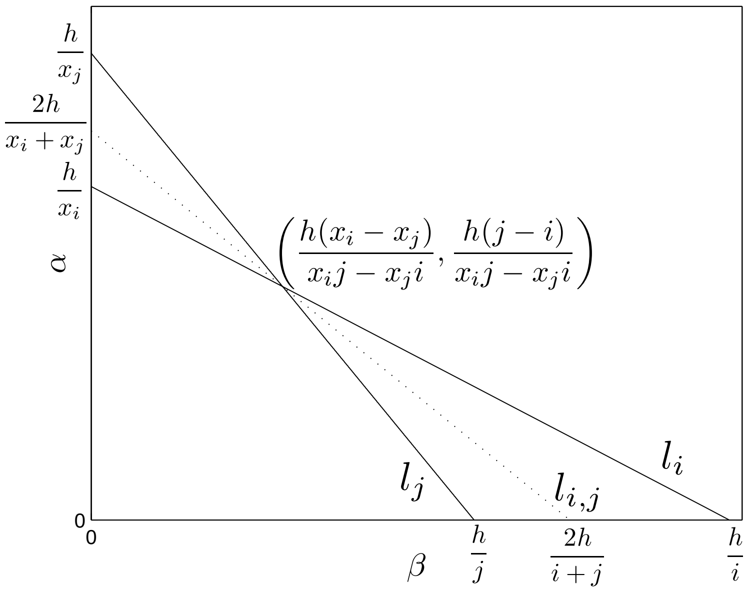


Figure 4.4: The lines l_i, $l_{i,j}$ and l_j.

falls 'below' the intersection of the open half spaces determined by l_i and l_j (sharing only a single point with the boundary) and therefore it must fall 'below' the convex set determined by $\alpha = F(\beta)$ (again sharing at most a single point with $\alpha = F(\beta)$). Therefore, by Proposition 4.2.3, we know the term can be ignored.

All of the cross terms fall 'below' the space determined by $\alpha = F(\beta)$, therefore the only terms that determine $\alpha = G(\beta)$ are those from the first summation. Therefore $F(\beta) = G(\beta)$ and $g_{\alpha,\beta,h}(n)$ has the same boundary as $f_{\alpha,\beta,h}(n)$. Hence by Proposition 4.2.2, $g_{\alpha,\beta,h}(n) = f^2_{\alpha,\beta,h}(n) \equiv f_{\alpha,\beta,h}(n)$.

QED.

4.3 Dividing Lines

Let $f_{\alpha,\beta,h_f}(n), g_{\alpha,\beta,h_g}(n) \in \mathcal{F}$, and let

$$R(f_{\alpha,\beta,h_f}(n), g_{\alpha,\beta,h_g}(n)) = \frac{f_{\alpha,\beta,h_f}(n)}{g_{\alpha,\beta,h_g}(n)} = \frac{\sum_{i=0}^{m-1} a_i n^{h_f - x_i\alpha - \beta i}}{\sum_{j=0}^{m-1} b_j n^{h_g - y_j\alpha - \beta j}}$$

be the ratio of the two posynomials. In Chapter 5, we will discuss examples of this type of ratio and will want to know for which values of β and α, do we have $g_{\alpha,\beta,h_g}(n) \to \infty$ and $R(f_{\alpha,\beta,h_f}(n), g_{\alpha,\beta,h_g}(n)) \to 0$, as $n \to \infty$. So in the remainder of this section, we assume that $g_{\alpha,\beta,h_g}(n) \to \infty$ as $n \to \infty$. Now, if $f_{\alpha,\beta,h_f}(n) \to 0$ as $n \to \infty$ or is constant, then clearly $R(f_{\alpha,\beta,h_f}(n), g_{\alpha,\beta,h_g}(n)) \to 0$. So, we also assume that $f_{\alpha,\beta,h_f}(n) \to \infty$ as $n \to \infty$.

We rewrite $R(f_{\alpha,\beta,h_f}(n), g_{\alpha,\beta,h_g}(n))$ as

$$R(f_{\alpha,\beta,h_f}(n), g_{\alpha,\beta,h_g}(n)) = \sum_{i=0}^{m-1} \frac{a_i n^{h_f - x_i\alpha - \beta i}}{g_{\alpha,\beta,h_g}(n)} = \sum_{i=0}^{m-1} \left(\frac{a_i n^{h_f - x_i\alpha - \beta i}}{\sum_{j=0}^{m-1} b_j n^{h_g - y_j\alpha - \beta j}} \right).$$

Then $R(f_{\alpha,\beta,h_f}(n), g_{\alpha,\beta,h_g}(n))$ goes to zero if and only if for each $0 \le i \le m-1$, either $a_i = 0$ or $\frac{a_i n^{h_f - x_i\alpha - \beta i}}{g_{\alpha,\beta,h_g}(n)} \to 0$. So, to answer our initial query, for each $0 \le i \le m-1$ with $a_i = 1$, we need to determine for which values of β and α do we have $\frac{a_i n^{h_f - x_i\alpha - \beta i}}{g_{\alpha,\beta,h_g}(n)} \to 0$ as $n \to \infty$.

Lemma 4.3.1. *Let $h_1, h_2, m \in \mathbb{Z}_+$. Let $i, j \in \{0, 1, \ldots, m-1\}$ and $a_i, b_j \in \{0, 1\}$. If* *for any* $j_0 \in \{0, 1, \ldots, m-1\}$

$$\frac{a_i n^{h_1 - x_i\alpha - \beta i}}{b_{j_0} n^{h_2 - y_{j_0}\alpha - \beta j_0}} \to 0,$$

then as $n \to \infty$

$$\frac{a_i n^{h_1 - x_i\alpha - \beta i}}{\sum_{j=0}^{m-1} b_j n^{h_2 - y_j\alpha - \beta j}} \to 0.$$

Proof:

Suppose that there exists $j_0 \in \{0, 1, \ldots, m-1\}$ such that $\frac{a_i n^{h_1 - x_i\alpha - \beta i}}{b_{j_0} n^{h_2 - y_{j_0}\alpha - \beta_{j_0}}} \to 0$ as $n \to \infty$. Then

$$0 \leq \frac{a_i n^{h_1 - x_i\alpha - \beta i}}{\sum_{j=0}^{m-1} b_j n^{h_2 - y_j\alpha - \beta j}} \leq \frac{a_i n^{h_1 - x_i\alpha - \beta i}}{b_{j_0} n^{h_2 - y_{j_0}\alpha - \beta_{j_0}}} \to 0.$$

QED.

So, to show that $\frac{a_i n^{h_f - x_i\alpha - \beta i}}{g_{\alpha,\beta,h_g}(n)} \to 0$ at (β, α), we only need to show that there exists some j_0 such that $\frac{a_i n^{h_f - x_i\alpha - \beta i}}{b_{j_0} n^{h_g - x_{j_0}\alpha - \beta_{j_0}}} \to 0$ at (β, α).

Now, as we discussed earlier in this chapter, each term of the form $a_i n^{h_f - x_i\alpha - \beta i}$ has an an associated line l_i that divides the β, α-parameter space. Likewise, each posynomial $g_{\alpha,\beta,h_g}(n) \in \mathcal{F}$ has an associated convex set $\mathcal{C}(g)$ that also divides the β, α-parameter space into regions with different limiting behaviors. We use these representations to determine when $\frac{a_i n^{h_f - x_i\alpha - \beta i}}{g_{\alpha,\beta,h_g}(n)} \to 0$ as $n \to \infty$ and to this end, we have the following definition and lemmas.

Definition 4.3.2. *Let* l_1 *and* l_2 *be lines and* $(\beta, \alpha) > (0, 0)$ *be a point in the parameter* *space. Let $l_{\beta,\alpha}$ be the unique line through the origin and (β, α). We say that l_1 is* <u>*under*</u> *l_2 at the point (β, α), if $l_{\beta,\alpha}$ intersects l_1 'below' l_2, i.e., the point of intersection between l_1 and $l_{\beta,\alpha}$ lies in the open half space defined by l_2 that contains the origin.*

Lemma 4.3.3. *Let $h_1, h_2, m \in \mathbb{Z}_+$ with $h_1 < h_2$. Let $i, j \in \{0, 1, \ldots, m-1\}$, $0 < x_i, y_j \leq m-1$, and $a_i = b_j = 1$. Let*

$$r_{\alpha,\beta,h_1,h_2}(n) = \frac{a_i n^{h_1 - x_i\alpha - \beta i}}{b_j n^{h_2 - y_j\alpha - \beta j}} = n^{(h_1-h_2)+(y_j-x_i)\alpha+(j-i)\beta}.$$

Let l_i be the line $h_1 - x_i\alpha - \beta i = 0$ and l_j be the line $h_2 - y_j\alpha - \beta j$. Then $r_{\alpha,\beta,h_1,h_2}(n) \to 0$ as $n \to \infty$ at all points (β, α) for which l_i is under l_j at (β, α).

Proof:

Let l_i be under l_j at (β^*, α^*). Therefore l_{β^*,α^*} intersects l_i 'below' l_j. Then letting α_i and α_j be the α-intercepts of l_i and l_j, respectively, clearly this can only happen in the following four ways:

- when $\alpha_i < \alpha_j$ and l_i and l_j intersect in the β, α-parameter space
- when $\alpha_i \geq \alpha_j$ and l_i and l_j intersect in the β, α-parameter space
- when $\alpha_i < \alpha_j$ and l_i and l_j do not intersect in the β, α-parameter space, but are not parallel
- when $\alpha_i < \alpha_j$ and l_i and l_j are parallel.

Note that $\alpha_i \geq \alpha_j$ need not be considered if the lines do not intersect in the β, α-parameter space because then l_i would not be 'below' l_j.

Next, note that

$$r_{\alpha,\beta,h_1,h_2}(n) = n^{(h_1-h_2)+(y_j-x_i)\alpha+(j-i)\beta} \longrightarrow 0$$

as $n \to \infty$, if and only if $(h_1 - h_2) + (y_j - x_i)\alpha + (j - i)\beta < 0$. Also, at the origin,

$$r_{0,0,h_1,h_2}(n) = n^{(h_1-h_2)+(y_j-x_i)0+(j-i)0} = n^{h_1-h_2} \longrightarrow 0$$

as $n \to \infty$ since $h_1 < h_2$. Let l_3 be the line

$$l_3 : (h_1 - h_2) + (y_j - x_i)\alpha + (j - i)\beta = 0.$$

Then we have that $r_{\alpha,\beta,h_1,h_2}(n) \longrightarrow 0$ in the open half space 'below' l_3. Hence it is sufficient to show that (β^*, α^*) falls within the open half space 'below' l_3.

Also note that if l_i and l_j are not parallel, then l_3 goes through their point of intersection since

$$h_1 - x_i\alpha - i\beta = h_2 - y_j\alpha - j\beta$$

implies that

$$(h_1 - h_2) + (y_j - x_i)\alpha + (j - i)\beta = 0.$$

The α-intercepts of l_i, l_j and l_3 are $\alpha_i = \frac{h_1}{x_i}, \alpha_j = \frac{h_2}{y_j}$ and $\alpha_3 = \frac{h_1 - h_2}{x_i - y_j}$, respectively. Also, the β-intercepts of l_i, l_j and l_3 are $\beta_i = \frac{h_1}{i}, \beta_j = \frac{h_2}{j}$ and $\beta_3 = \frac{h_1 - h_2}{i - j}$, respectively. Finally, the slopes of l_i, l_j and l_3 are $s_1 = -i/x_i, s_2 = -j/y_j$ and $s_3 = \frac{i-j}{y_j - x_i}$, respectively.

The reminder proof is divided into the four possible cases mentioned earlier.

Case 1: $\alpha_i < \alpha_j$ and l_i and l_j intersect in the β, α-parameter space.

First note that $\alpha_i < \alpha_j$ and l_i and l_j intersect in the β, α-parameter space, imply that $\beta_j \leq \beta_i$.

Let $(\beta_{int}, \alpha_{int})$ be the point where l_i and l_j intersect. Let A be the closed triangle formed by the α-axis, l_i, and the line $l_{\beta_{int}, \alpha_{int}}$ minus the point $(\beta_{int}, \alpha_{int})$. Then clearly $(\beta^*, \alpha^*) \in A$.

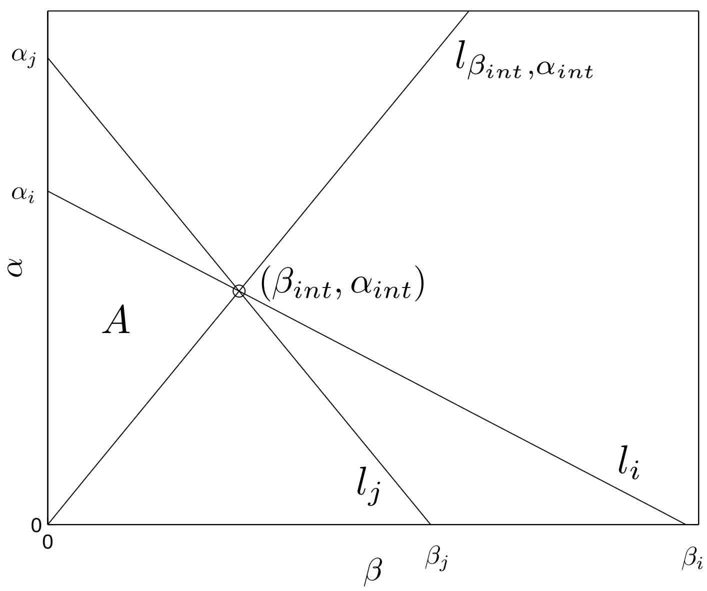


Figure 4.5: Case 1: The region A.

Now, $\beta_i = \frac{h_1}{i} \geq \beta_j = \frac{h_2}{j}$ and $h_1 < h_2$ together imply that $j > i$. Therefore, we have that the β-intercept of l_3 is $\beta_3 = \frac{h_1 - h_2}{i-j} > 0$. Also, recall that l_3 goes through the point $(\beta_{int}, \alpha_{int})$. If $x_i > y_j$, then the slope of l_3 is positive and all of A falls 'below' l_3. If $x_i = y_j$ then l_3 is a vertical line and again all of A falls 'below' l_3. Finally, if $x_i < y_j$ then the slope of l_3 is negative and we must check that the α-intercept of l_3 does not fall 'below' the α-intercept of l_i. Now, $\alpha_j > \alpha_i$, i.e.,

$$\frac{h_2}{y_j} > \frac{h_1}{x_i}$$

and we have that

$$h_2 x_i > h_1 y_j$$

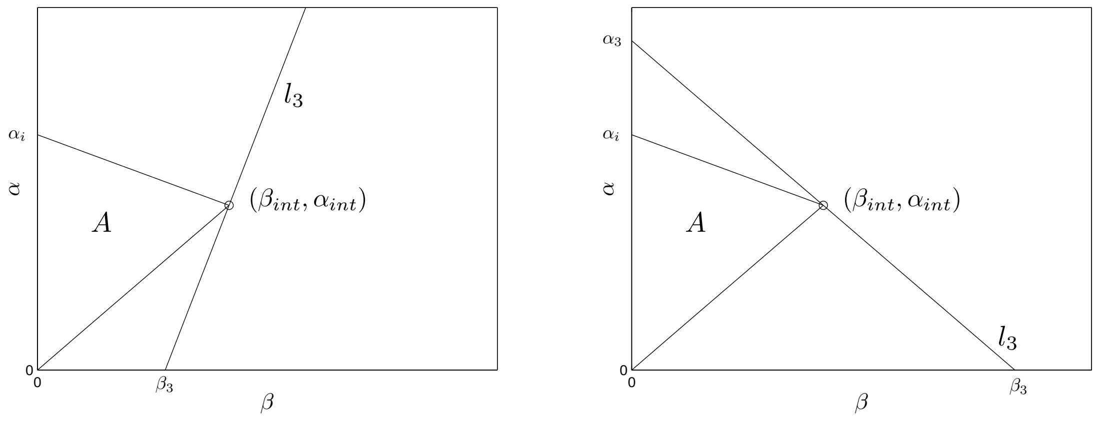
$$h_2 x_i - h_1 x_i > h_1 y_j - h_1 x_i$$

$$x_i(h_2 - h_1) > h_1(y_j - x_i)$$

(noting that $x_i < y_j$ implies that $y_j - x_i > 0$)

$$\frac{h_2 - h_1}{y_j - x_i} > \frac{h_1}{x_i}.$$

Therefore $x_i < y_j$ implies that $\alpha_3 > \alpha_i$ and we again have that all of A falls 'below' l_3. Therefore, since $(\beta^*, \alpha^*) \in A$, we have that $(h_1 - h_2) + (y_j - x_i)\alpha^* + (j - i)\beta^* < 0$ and hence $r_{\alpha^*, \beta^*, h_1, h_2}(n) \to 0$ as $n \to \infty$.



(a) **(b)**

Figure 4.6: Case 1 with **(a)** $x_i > y_j$ and the slope of l_3 positive. **(b)** $x_i < y_j$ and the slope of l_3 negative.

Case 2: $\alpha_i \geq \alpha_j$ and l_i and l_j intersect in the β, α-parameter space.

First note that $\alpha_i \geq \alpha_j$ and l_i and l_j intersect in the β, α-parameter space, imply that $\beta_j > \beta_i$.

Let $(\beta_{int}, \alpha_{int})$ be the point where l_i and l_j intersect. Let A' be the closed triangle formed by the β-axis, l_i, and the line $l_{\beta_{int},\alpha_{int}}$ minus the point $(\beta_{int}, \alpha_{int})$. Then clearly $(\beta^*, \alpha^*) \in A'$.

Then by an argument similar to that of Case 1, we have that A' always falls 'below' l_3. Therefore, since $(\beta^*, \alpha^*) \in A'$, we have that $(h_1 - h_2) + (y_j - x_i)\alpha^* + (j - i)\beta^* < 0$ and hence $r_{\alpha^*,\beta^*,h_1,h_2}(n) \to 0$ as $n \to \infty$.

Case 3: $\alpha_i < \alpha_j$ and l_i and l_j do not intersect in the β, α-parameter space, but are not parallel.

First note that $\alpha_i < \alpha_j$ and l_i and l_j do not intersect in the β, α-parameter space,

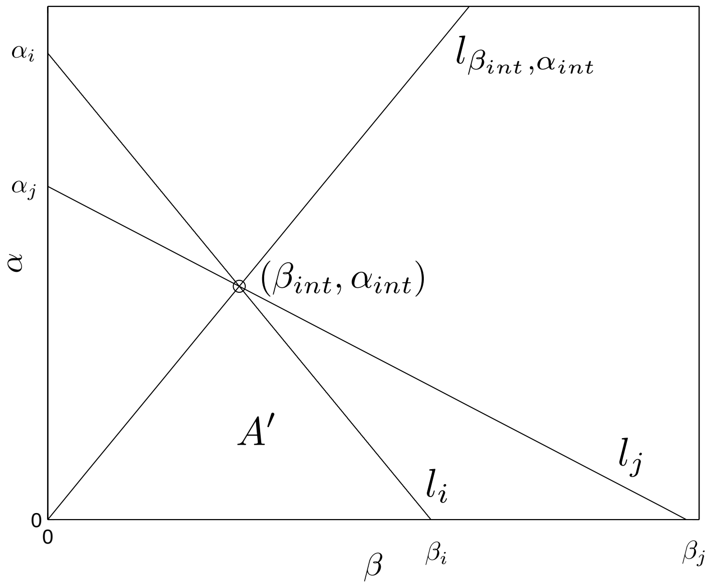


Figure 4.7: Case 2: The region A'.

imply that $\beta_i < \beta_j$ and the entire line l_i falls 'below' l_j within the β, α-parameter space. In this instance we consider the closed triangle A'' formed by the α- and β-axes and l_i. Then clearly $(\beta^*, \alpha^*) \in A''$.

Let $(\beta_{int}, \alpha_{int})$ be the point where l_i and l_j intersect. Recall that l_3 goes through the point $(\beta_{int}, \alpha_{int})$.

First let us assume that $\beta_{int} < 0$. Then $\alpha_{int} > 0$ since both l_i and l_j have negative slopes. Then if $x_i > y_j$, we see that the α-intercept of l_3 is negative and so the entire β, α-parameter space falls 'below' l_3. Hence, A'' falls 'below' l_3. Also, if $x_i = y_j$, then l_3 is a vertical line and again all of A'' falls 'below' l_3. If $x_i < y_j$ then $\alpha_3 > \alpha_i$ by the same argument used in Case 1. If $i > j$, then $\beta_3 < 0$ and A'' is 'below' l_3. Also if $i = j$, then l_3 is a horizontal line and again A'' is 'below' l_3. Finally, consider the case

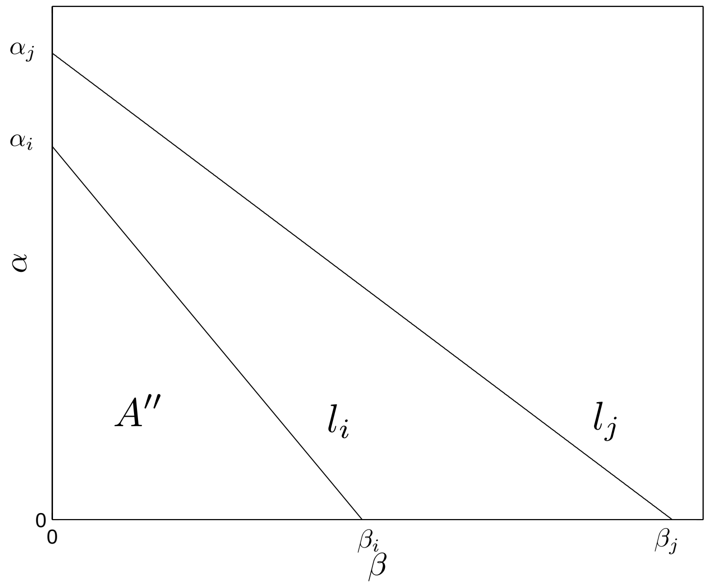


Figure 4.8: Case 3: The region A''.

when $x_i < y_j$ and $i \leq j$. Now, $\beta_j > \beta_i$, i.e.,

$$\frac{h_2}{j} > \frac{h_1}{i}$$

and we have that

$$h_2\, i > h_1\, j$$

$$h_2\, i - h_1\, i > h_1\, j - h_1\, i$$

$$i\,(h_2 - h_1) > h_1\,(j - i)$$

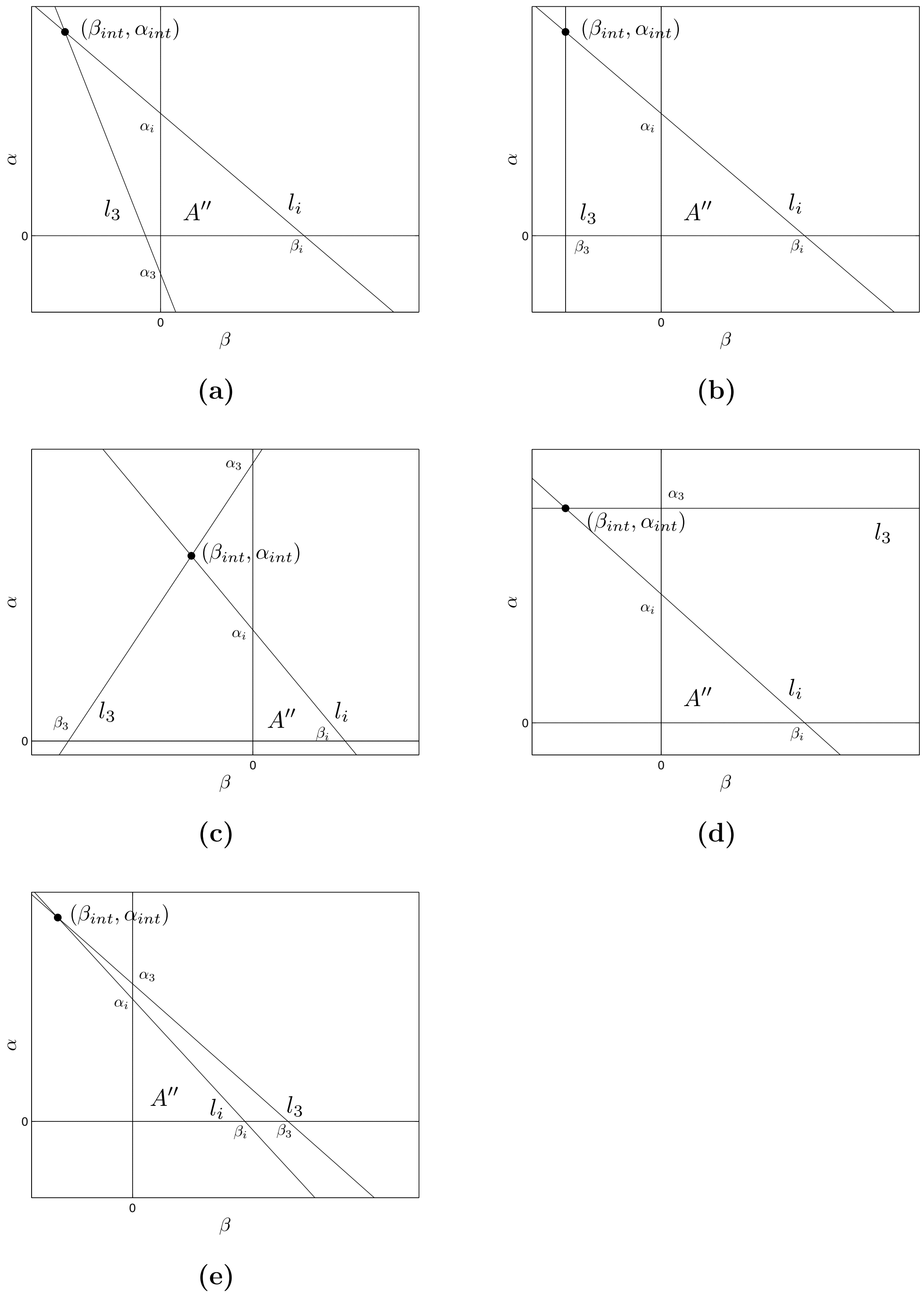


Figure 4.9: Case 3 with $\beta_{int} < 0$ and **(a)** $x_i > y_j$. **(b)** $x_i = y_j$. **(c)** $x_i < y_j$ and $i < j$. **(d)** $x_i < y_j$ and $i = j$. **(e)** $x_i < y_j$ and $i > j$.

(noting that $i < j$ implies that $j - i > 0$)

$$\frac{h_2 - h_1}{j - i} > \frac{h_1}{i}.$$

Therefore $i < j$ implies that $\beta_3 > \beta_i$. So, $\alpha_3 > \alpha_i$ and $\beta_3 > \beta_i$, thus l_i falls 'below' l_3 and hence A'' is 'below' l_3. Therefore, since $(\beta^*, \alpha^*) \in A''$, we have that $(h_1 - h_2) + (y_j - x_i)\alpha^* + (j - i)\beta^* < 0$ and $r_{\alpha^*,\beta^*,h_1,h_2}(n) \to 0$ as $n \to \infty$ whenever $\beta_{int} < 0$.

Now, assume that $\beta_{int} > 0$. (Note that we do not have to consider the case when $\beta_{int} = 0$ since that is covered by Case 1.) Then $\alpha_{int} < 0$ since both l_i and l_j have negative slopes. Then by an argument similar to that above (switching α and β) we have that since $(\beta^*, \alpha^*) \in A''$, $r_{\alpha^*,\beta^*,h_1,h_2}(n) \to 0$ as $n \to \infty$ whenever $\beta_{int} > 0$.

So, we see that $(\beta^*, \alpha^*) \in A''$ gives $r_{\alpha^*,\beta^*,h_1,h_2}(n) \to 0$ as $n \to \infty$ whenever $\alpha_i < \alpha_j$ and l_i and l_j do not intersect in the β, α-parameter space, but are not parallel.

Case 4: $\alpha_i < \alpha_j$ and l_i and l_j are parallel.

In this instance, we again consider points in the closed triangle A'' formed by the α- and β-axes and l_i. Then clearly $(\beta^*, \alpha^*) \in A''$.

Since l_i and l_j are parallel, we have that their slopes are equal and thus, $\frac{i}{x_i} = \frac{j}{y_j}$. Now, $i, j > 0$ therefore there exists $c > 0$ such that $j = c\,i$ and hence $y_j = c\,x_i$. We can rewrite $r_{\alpha,\beta,h_1,h_2}(n)$ as

$$r_{\alpha,\beta,h_1,h_2}(n) = n^{(h_1-h_2)+(c-1)x_i\alpha+(c-1)i\beta}.$$

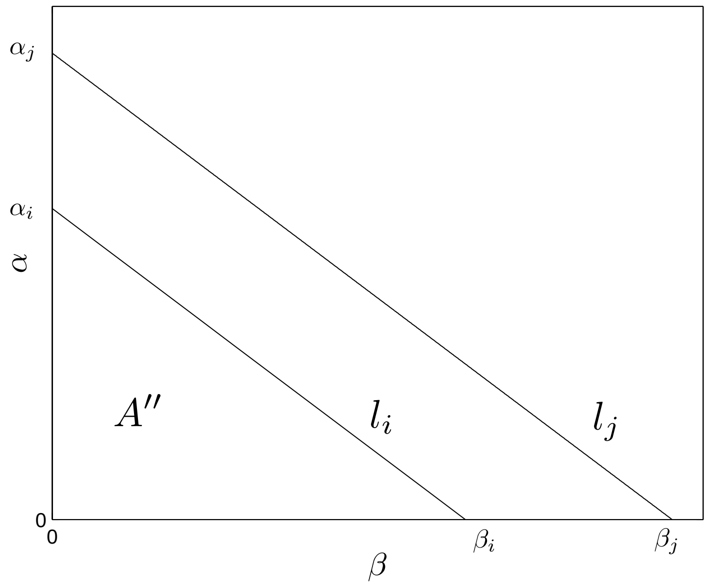


Figure 4.10: Case 4: The region A''.

Also, l_3 has slope $s_3 = \frac{-(c-1)i}{(c-1)x_i} = \frac{-i}{x_i}$ and is parallel to l_i and l_j.

If $c \leq 1$, then $(c-1) \leq 0$ and this along with $h_1 < h_2$ implies that $(h_1 - h_2) + (c-1)x_i\alpha + (c-1)i\beta < 0$. Hence $r_{\alpha,\beta,h_1,h_2}(n) \to 0$ as $n \to \infty$ for all $\alpha, \beta \geq 0$. If $c > 1$ then we see that $y_j > x_i$. As in Case 1, $y_j > x_i$ and $\alpha_i < \alpha_j$ together imply that $\alpha_3 > \alpha_i$. However, l_3 is parallel to l_i and has a greater α-intercept. Therefore, l_i, and hence A'', falls in the open half space 'below' l_3. Hence, $r_{\alpha^*,\beta^*,h_1,h_2}(n) \to 0$ as $n \to \infty$ since $(\beta^*, \alpha^*) \in A''$.

So, $(\beta^*, \alpha^*) \in A''$ gives $r_{\alpha^*,\beta^*,h_1,h_2}(n) \to 0$ as $n \to \infty$ whenever $\alpha_i < \alpha_j$ and l_i and l_j are parallel.

Therefore whenever l_i is under l_j at (β, α), we have that $r_{\alpha,\beta,h_1,h_2}(n) \to 0$ as $n \to \infty$.

QED.

We now generalize this notion of a line being under implying convergence to zero to ratios of the form $\frac{a_i n^{h_1 - x_i\alpha - \beta i}}{g_{\alpha,\beta,h_g}(n)}$. It is indeed true that $\frac{a_i n^{h_1 - x_i\alpha - \beta i}}{g_{\alpha,\beta,h_g}(n)} \to 0$ as $n \to \infty$, whenever l_i is completely under the convex set $\mathcal{C}(g)$ in the obvious sense.

Lemma 4.3.4. *Let $h_1, h_g, m \in \mathbb{Z}_+$ with $h_1 < h_g$. Let $i, j \in \{0, 1, \ldots, m-1\}$, $0 < x_i, y_j \leq m-1$, and $a_i = 1$ and $b_j \in \{0, 1\}$ not all zero. Let*

$$r_{\alpha,\beta,h_1,g}(n) = \frac{a_i n^{h_1 - x_i\alpha - \beta i}}{g_{\alpha,\beta,h_g}(n)} = \frac{a_i n^{h_1 - x_i\alpha - \beta i}}{\sum_{j=0}^{m-1} b_j n^{h_g - y_j\alpha - \beta j}}.$$

Let l_i be the line $h_1 - x_i\alpha - \beta i = 0$ and $\mathcal{C}(g)$ be the convex region in which $g_{\alpha,\beta,h_g}(n) \to 0$ as described in Proposition 4.1.1. Then $r_{\alpha,\beta,h_1,g}(n) \to 0$ whenever $g_{\alpha,\beta,h_g}(n) \to \infty$ (as $n \to \infty$) and l_i falls 'below' $\mathcal{C}(g)$.

Proof:

Let l_i fall 'below' $\mathcal{C}(g)$. Let (β, α) be a point in the β, α-parameter space for which $g_{\alpha,\beta,h_g}(n) \to \infty$ as $n \to \infty$. Then, (β, α) also falls in the region 'below' $\mathcal{C}(g)$.

If (β, α) is 'above' the line l_i, then $a_i n^{h_1 - x_i\alpha - \beta i} \to 0$ as $n \to \infty$, and clearly, $r_{\alpha,\beta,h_1,g}(n) \to 0$. Also, if (β, α) is on l_i, then $h_1 - x_i\alpha - \beta i = 0$ and again $r_{\alpha,\beta,h_1,g}(n) \to 0$ as $n \to \infty$. So, from now on, we assume that (β, α) is in the region 'below' l_i.

Now, (β, α) is 'below' l_i which is 'below' $\mathcal{C}(g)$, therefore the line $l_{\beta,\alpha}$ will intersect l_i 'below' $\mathcal{C}(g)$. Let $(\beta_{int}, \alpha_{int})$ be the point where $l_{\beta,\alpha}$ intersects $G(\beta)$, the boundary of $\mathcal{C}(g)$. Recall that $G(\beta)$ is a piecewise linear function, whose pieces come from the lines associated with the terms in $g_{\alpha,\beta,h_g}(n)$. Therefore, there must exist some $0 \leq j \leq m-1$ for which $b_j = 1$ and $(\beta_{int}, \alpha_{int})$ is on the line l_j. Therefore, l_i is under l_j at the point (β, α). Hence, by Lemma 4.3.3, we have that

$$\frac{a_i n^{h_1 - x_i\alpha - \beta\, i}}{b_j n^{h_g - y_j\alpha - \beta\, j}} \to 0$$

as $n \to \infty$. Therefore by Lemma 4.3.1 we have that all of

$$r_{\alpha,\beta,h_1,g}(n) = \frac{a_i n^{h_1 - x_i\alpha - \beta\, i}}{g_{\alpha,\beta,h_g}(n)} = \frac{a_i n^{h_1 - x_i\alpha - \beta\, i}}{\sum_{j=0}^{m-1} b_j n^{h_g - y_j\alpha - \beta j}} \to 0$$

as $n \to \infty$.

QED.

So, we have that whenever line l_i falls in the open region 'below' $\mathcal{C}(g)$ we have that the associated term over $g_{\alpha,\beta,h_g}(n)$ goes to 0.

We are now ready for the main theorem of this section.

Theorem 4.3.5. *Let $f_{\alpha,\beta,h_f}(n), g_{\alpha,\beta,h_g}(n) \in \mathcal{F}$. Let $h_f < h_g$ and*

$$R(f_{\alpha,\beta,h_f}(n), g_{\alpha,\beta,h_g}(n)) = \frac{f_{\alpha,\beta,h_f}(n)}{g_{\alpha,\beta,h_g}(n)} = \frac{\sum_{i=0}^{m-1} a_i n^{h_f - x_i\alpha - \beta i}}{\sum_{j=0}^{m-1} b_j n^{h_g - x_j\alpha - \beta j}}.$$

Then $R(f_{\alpha,\beta,h_f}(n), g_{\alpha,\beta,h_g}(n)) \to 0$ whenever $g_{\alpha,\beta,h_g}(n) \to \infty$ as $n \to \infty$, if for each $0 \leq i \leq m-1$, either $a_i = 0$ or l_i is 'below' $\mathcal{C}(g)$.

Proof:

First note that $R(f_{\alpha,\beta,h_f}(n), g_{\alpha,\beta,h_g}(n))$ will only go to zero if and only if for each $0 \leq i \leq m-1$, $\frac{a_i n^{h_1 - x_i\alpha - \beta\, i}}{g_{\alpha,\beta,h_g}(n)} \to 0$.

If $a_i = 0$, then clearly $\frac{a_i n^{h_1 - x_i\alpha - \beta\, i}}{g_{\alpha,\beta,h_g}(n)} \to 0$. So, let $a_i = 1$ and l_i be 'below' $\mathcal{C}(g)$. Then by Lemma 4.3.4, we have that $\frac{a_i n^{h_1 - x_i\alpha - \beta\, i}}{g_{\alpha,\beta,h_g}(n)} \to 0$ and hence $R(f_{\alpha,\beta,h_f}(n), g_{\alpha,\beta,h_g}(n)) \to 0$ as $n \to \infty$.

QED.

So, to show that $R(f_{\alpha,\beta,h_f}(n), g_{\alpha,\beta,h_g}(n))$ goes to zero, it is sufficient to show that each of the lines associated with $f_{\alpha,\beta,h_f}(n)$ falls 'below' the convex set $\mathcal{C}(g)$. This result gives us the following obvious corollary relating the two convex sets $\mathcal{C}(f)$ and $\mathcal{C}(g)$.

Corollary 4.3.6. *Let $f_{\alpha,\beta,h_f}(n), g_{\alpha,\beta,h_g}(n) \in \mathcal{F}$. Let $h_f < h_g$ and*

$$R(f_{\alpha,\beta,h_f}(n), g_{\alpha,\beta,h_g}(n)) = \frac{f_{\alpha,\beta,h_f}(n)}{g_{\alpha,\beta,h_g}(n)}$$

Then $R(f_{\alpha,\beta,h_f}(n), g_{\alpha,\beta,h_g}(n)) \to 0$ whenever $g_{\alpha,\beta,h_g}(n) \to \infty$ as $n \to \infty$, if $F(\beta)$ falls in the region 'below' $\mathcal{C}(g)$, i.e., if the region $\mathcal{C}(f)$ contains the region $\mathcal{C}(g)$.

Chapter 5

The Discrete Model

In this chapter we present the Discrete Random Dot Product Graph. We begin by introducing the model and proving basic results in Section 5.1.1. In Section 5.1.2 we illustrate the difficulty of calculations in this model by finding the threshold for the appearance of K_3 as a subgraph. In Section 5.2, we define the *probability order polynomial*, or POP, of a graph H, give a general method for calculating the POP of H, and present formulas for the POPs of trees, cycles, and complete graphs. In Section 5.3, we present first moment results for trees, cycles, and complete graphs. We also prove a threshold result for K_3 and describe a general method for proving threshold results when all the required POPs are known.

5.1 The 0, 1 Discrete Random Dot Product Graph

In this chapter we study the behavior of the Random Dot Product Graph when the vectors are drawn from the discrete sample space $\{0,1\}^t$, where $t \in \mathbb{Z}_{\geq 1}$. In this case, each vertex $v \in V(G)$ is given a vector x_v drawn from $\{0,1\}^t$ as follows: each coordinate of x_v is independently assigned the value 1 with probability p and the value 0 with probability $1-p$, where $p \in [0,1]$. We define the probability mapping $f : \mathbb{R} \to [0,1]$ to be $f(r) = r/t$, guaranteeing that the dot product of any two vectors is mapped into probabilities. We denote this sample space of discrete random dot product graphs as $\mathcal{D}[n, \{0,1\}^t, p]$.

In [20], Singer introduces Random Intersections Graphs, a random graph model in which vertices are mapped to random sets in the following way: each vertex v is assigned a subset S_v of a universal set of t elements, $[t]$, by independently placing each $i \in [t]$ in S_v with probability p. If we view the vector x_v as an indicator function in which element i is placed in S_v if and only if the ith coordinate of x_v is one, then the assignment of vectors to vertices in a Discrete Random Dot Product Graph is similar to Singer's assignment of sets to vertices. However, unlike in a Discrete Random Dot Product Graph, in a Random Intersection Graph, once the sets have been assigned, the existence of edges is not random, but deterministic, with an edge between vertices u and v if and only if $S_u \cap S_v \neq \emptyset$. Although the Random Intersection Graph is different from the discrete model we present in this Chapter, it is a Random Dot Product Graph model in which the probability mapping $f(y)$ is 0 when $y = 0$

and 1 otherwise.

5.1.1 Basic Results

Proposition 5.1.1. *Let G be drawn from $\mathcal{D}[n, \{0,1\}^t, p]$. For any $u, v \in V(G)$ we have $P[u \sim v] = p^2$.*

Proof:

Let $u, v \in V(G)$. Let $\mathbf{x} = [x_1, x_2, \ldots, x_t]^T$ and $\mathbf{y} = [y_1, y_2, \ldots, y_t]^T$ be the vectors of u and v respectively. Then

$$P[u \sim v] = E[\frac{\mathbf{x} \cdot \mathbf{y}}{t}] = E[\frac{(x_1 y_1 + \cdots + x_t y_t)}{t}] = \frac{tE[x_1 y_1]}{t} = E[x_1]E[y_1] = p^2.$$

QED.

Thus, the probability of an edge is p^2 and the expected number of edges is $\binom{n}{2}p^2$. While, this constant probability on the edges is also common to Erdős-Rényi random graphs, our graphs are not simply Erdős-Rényi (n, p^2). For example in an Erdős-Rényi (n, p^2) random graph the probability that two vertices are adjacent given that share a common neighbor is again p^2. However in this model, the knowledge of the common neighbor increases the probability of the adjacency, i.e., the model exhibits clustering.

Lemma 5.1.2. *Let G be drawn from $\mathcal{D}[n, \{0,1\}^t, p]$. For any $u, v, w \in V(G)$, we* *have*

$$P[u \sim v \sim w] = \frac{1}{t^2}[tp^3 + t(t-1)p^4].$$

Proof:

Let $u, v, w \in V(G)$. Let $\mathbf{x} = [x_1, x_2, \ldots, x_t]^T$, $\mathbf{y} = [y_1, y_2, \ldots, y_t]^T$, and $\mathbf{z} = [z_1, z_2, \ldots, z_t]^T$, be the vectors of u,v, and w respectively. Then

$$P[u \sim v \sim w] = E\left[\left(\frac{\mathbf{x}\cdot\mathbf{y}}{t}\right)\left(\frac{\mathbf{y}\cdot\mathbf{z}}{t}\right)\right] = E\left[\left(\frac{x_1y_1 + \cdots + x_ty_t}{t}\right)\left(\frac{y_1z_1 + \cdots + y_tz_t}{t}\right)\right]$$

$$= \frac{1}{t^2}\sum_{i,j\in\{1,2,\ldots,t\}} E\left[x_iy_iy_jz_j\right].$$

We now consider what occurs as i and j vary. If $i = j$, then the term inside the sum is $E[x_iy_iy_jz_j] = E[x_iy_i^2z_i] = E[x_iy_iz_i]$, since $y_i \in \{0, 1\}$. Furthermore, $E[x_iy_iz_i] = E[x_i]E[y_i]E[z_i] = p^3$, since x_i, y_i and z_i are independent. Also, if $i \neq j$, then y_i and y_j are independent and so the term inside the sum is $E[x_iy_iy_jz_j] = E[x_i]E[y_i]E[y_j]E[z_j] = p^4$. Therefore we see that

$$P[u \sim v \sim w] = \frac{1}{t^2}\left(\sum_{i\in\{1,2,\ldots,t\}} p^3 + \sum_{i,j\in\{1,2,\ldots,t\}\,:\,i\neq j} p^4\right) = \frac{1}{t^2}\left(tp^3 + t(t-1)p^4\right).$$

QED.

Similarly and as will be shown in Section 5.1.2, we obtain the following result.

Lemma 5.1.3. *Let G be drawn from $\mathcal{D}[n, \{0,1\}^t, p]$. For any $u, v, w \in V(G)$ we have*

$$P[\text{the triangle } uvw] = \frac{1}{t^3}[tp^3 + 3t(t-1)p^5 + t(t-1)(t-2)p^6].$$

These two lemmas give us the following result.

Proposition 5.1.4. *Let G be drawn from $\mathcal{D}[n, \{0,1\}^t, p]$. For any $u, v, w \in V(G)$ we have $P[u \sim w | u \sim v \sim w] \geq P[u \sim w]$, with equality holding iff $p = 1$.*

Proof:

$$P[u \sim w | u \sim v \sim w] = \frac{P[\text{the triangle } uvw]}{P[u \sim v \sim w]}$$

which from Lemmas 5.1.3 and 5.1.2

$$= \frac{\frac{1}{t^3}[tp^3 + 3t(t-1)p^5 + t(t-1)(t-2)p^6]}{\frac{1}{t^2}[tp^3 + t(t-1)p^4]}$$

$$= \frac{1 + 3p^2(t-1) + p^3(t-1)(t-2)}{(1 + p(t-1))t}.$$

So,

$$P[u \sim w | u \sim v \sim w] - p^2 = \frac{(1-p)(1 + p + 2p^2(t-1))}{1 + p(-1+t))t} \geq 0$$

since $0 \leq p \leq 1$ and $t \geq 1$ implies that $1 - p \geq 0$, $1 + p + 2p^2(t-1) > 0$ and

$1 + p(t-1))t > 0$. Hence we have that

$$P[u \sim w | u \sim v \sim w] \geq p^2 = P[u \sim w]$$

with equality holding iff $p = 1$.

QED.

We have already determined that the probability of an edge is p^2 and therefore the expected number of edges is $\binom{n}{2}p^2$. However we would like to know for which values of p does a Discrete Random Dot Product Graph have edges (with high probability). Now, to further this discussion we must first calculate the variance on the number of edges in a Discrete Random Dot Product Graph.

Lemma 5.1.5. *Let G be drawn from $\mathcal{D}[n, \{0,1\}^t, p]$. Let the random variable $\mathbf{X}$ be the number of edges in G. Then*

$$Var[\mathbf{X}] = \binom{n}{2}p^2 + 6\binom{n}{3}\frac{p^3 + (t-1)p^4}{t} + \binom{n}{2}\binom{n-2}{2}p^4 - \binom{n}{2}^2 p^4.$$

Proof:

Consider $\mathbf{X}$ as the sum of the indicator functions for each individual edge. So, $\mathbf{X} = \sum_{\substack{u,v \in V(G) \\ u<v}} \mathbf{I}\{u \sim v\}$. Therefore by Proposition 5.1.1 we have that

$$E[\mathbf{X}] = E\left[\sum_{\substack{u,v \in V(G) \\ u<v}} \mathbf{I}\{u \sim v\}\right] = \sum_{\substack{u,v \in V(G) \\ u<v}} E[\mathbf{I}\{u \sim v\}] = \sum_{\substack{u,v \in V(G) \\ u<v}} P[u \sim v] = \binom{n}{2}p^2.$$

The variance of $\mathbf{X}$ is

$$\mathrm{Var}[\mathbf{X}] = E[\mathbf{X}^2] - E[\mathbf{X}]^2 = E[\mathbf{X}^2] - \binom{n}{2}^2 p^4$$

and so we calculate $E[\mathbf{X}^2]$.

$$E[\mathbf{X}^2] = E\left[\left(\sum_{u,v \in V(G)} \mathbf{I}\{u \sim v\}\right)^2\right]$$

$$= E\left[\sum_{\substack{u,v \in V(G) \\ u<v}} \mathbf{I}^2\{u \sim v\} + \sum_{\substack{u,v,w \in V(G) \\ u \neq v, u \neq w, v \neq w}} \mathbf{I}\{u \sim v\}\mathbf{I}\{v \sim w\}\right.$$

$$\left. + \sum_{\substack{u,v,w,z \in V(G) \\ u<v, w<z \\ \{u,v\} \cap \{w,z\} = \emptyset}} \mathbf{I}\{u \sim v\}\mathbf{I}\{w \sim z\}\right]$$

$$= E\left[\sum_{\substack{u,v \in V(G) \\ u<v}} \mathbf{I}^2\{u \sim v\}\right] + E\left[\sum_{\substack{u,v,w \in V(G) \\ u \neq v, u \neq w, v \neq w}} \mathbf{I}\{u \sim v\}\mathbf{I}\{v \sim w\}\right]$$

$$+ E\left[\sum_{\substack{u,v,w,z \in V(G) \\ u<v, w<z \\ \{u,v\} \cap \{w,z\} = \emptyset}} \mathbf{I}\{u \sim v\}\mathbf{I}\{w \sim z\}\right].$$

We now look at each of the three expected values individually. In the first expected value, since the indicator random variables can only have the value of 0 or 1 we see that

$$E\left[\sum_{\substack{u,v\in V(G)\\u<v}} \mathbf{I}^2\{u\sim v\}\right] = E\left[\sum_{\substack{u,v\in V(G)\\u<v}} \mathbf{I}\{u\sim v\}\right] = \sum_{\substack{u,v\in V(G)\\u<v}} E\left[\mathbf{I}\{u\sim v\}\right]$$

$$= \binom{n}{2} P[u\sim v] = \binom{n}{2} p^2.$$

Next, for the middle term, we have that

$$E\left[\sum_{\substack{u,v,w\in V(G)\\u\neq v,u\neq w,v\neq w}} \mathbf{I}\{u\sim v\}\,\mathbf{I}\{v\sim w\}\right] = \sum_{\substack{u,v,w\in V(G)\\u\neq v,u\neq w,v\neq w}} E\left[\mathbf{I}\{u\sim v\}\,\mathbf{I}\{v\sim w\}\right]$$

$$= 6\binom{n}{3} E\left[\mathbf{I}\{u\sim v\}\,\mathbf{I}\{v\sim w\}\right].$$

Now, $\mathbf{I}\{u\sim v\}\,\mathbf{I}\{v\sim w\} = 1$ if and only if $\mathbf{I}\{u\sim v\} = 1$ and $\mathbf{I}\{v\sim w\} = 1$, and is 0 otherwise. Therefore, $E\left[\mathbf{I}\{u\sim v\}\,\mathbf{I}\{v\sim w\}\right] = P[u\sim v\sim w] = \frac{p^3+(t-1)p^4}{t}$ by Lemma 5.1.2. Hence the middle term is

$$E\left[\sum_{\substack{u,v,w\in V(G)\\u\neq v,u\neq w,v\neq w}} \mathbf{I}\{u\sim v\}\,\mathbf{I}\{v\sim w\}\right] = 6\binom{n}{3}\frac{p^3+(t-1)p^4}{t}.$$

Finally, since the events $\{u\sim v\}$ and $\{w\sim z\}$ are independent, the last term becomes

$$E\left[\sum_{\substack{u,v,w,z\in V(G)\\ u<v,w<z\\ \{u,v\}\cap\{w,z\}=\emptyset}} \mathbf{I}\{u\sim v\}\,\mathbf{I}\{w\sim z\}\right] = \sum_{\substack{u,v,w,z\in V(G)\\ u<v,w<z\\ \{u,v\}\cap\{w,z\}=\emptyset}} E\left[\mathbf{I}\{u\sim v\}\right] E\left[\mathbf{I}\{w\sim z\}\right]$$

$$= \binom{n}{2}\binom{n-2}{2} P[u\sim v]P[w\sim z] = \binom{n}{2}\binom{n-2}{2} p^4.$$

Now we have that $E[\mathbf{X}^2] = \binom{n}{2}p^2 + 6\binom{n}{3}\frac{p^3+(t-1)p^4}{t} + \binom{n}{2}\binom{n-2}{2}p^4$. Therefore the variance is indeed

$$\mathrm{Var}[\mathbf{X}] = \binom{n}{2}p^2 + 6\binom{n}{3}\frac{p^3+(t-1)p^4}{t} + \binom{n}{2}\binom{n-2}{2}p^4 - \binom{n}{2}^2 p^4.$$

QED.

Theorem 5.1.6. *Let G be drawn from $\mathcal{D}[n,\{0,1\}^t,p]$. A threshold function for the appearance of edges in G is $p = 1/n$. That is, with high probability G will be edgeless whenever $p \ll \frac{1}{n}$ and with high probability G will have at least one edge whenever $p \gg \frac{1}{n}$, irrespective of t.*

Proof:

Let the random variable $\mathbf{X}$ be the number of edges in G. Then we know from the proof of Lemma 5.1.5 that $E(\mathbf{X}) = \binom{n}{2}p^2$. Let $1/n$ be our candidate for the threshold function for the appearance of edges. Now, by Markov's inequality $P[\mathbf{X} \geq 1] \leq E[\mathbf{X}] = \binom{n}{2}p^2$. So whenever the probability $p(n)$ is such that $p(n)/(1/n) \to 0$

as $n \to \infty$, we have that $E[\mathbf{X}] \to 0$ and therefore the $P[\mathbf{X} = 0] \to 1$ as $n \to \infty$ and almost every graph has no edges.

Also, whenever, $p(n)/(1/n) \to \infty$ as $n \to \infty$, then $E[\mathbf{X}] \to \infty$ and we use the second moment method. By Chebychev's inequality we know that $P[\mathbf{X} = 0] \leq \mathrm{Var}(\mathbf{X})/E[\mathbf{X}]^2$. By Lemma 5.1.5 we have that

$$P[\mathbf{X} = 0] \leq \frac{\mathrm{Var}[\mathbf{X}]}{E[\mathbf{X}]^2} = \frac{\binom{n}{2}p^2 + 6\binom{n}{3}\frac{p^3+(t-1)p^4}{t} + \binom{n}{2}\binom{n-2}{2}p^4 - \binom{n}{2}^2 p^4}{\left(\binom{n}{2}p^2\right)^2}$$

$$= \frac{\binom{n}{2}p^2}{\left(\binom{n}{2}p^2\right)^2} + \frac{6\binom{n}{3}\frac{p^3+(t-1)p^4}{t}}{\left(\binom{n}{2}p^2\right)^2} + \frac{\binom{n}{2}\binom{n-2}{2}p^4 - \binom{n}{2}^2 p^4}{\left(\binom{n}{2}p^2\right)^2}.$$

Now let us examine the three parts of the sum individually. First,

$$\frac{\binom{n}{2}p^2}{\left(\binom{n}{2}p^2\right)^2} = \frac{E[\mathbf{X}]}{E[\mathbf{X}]^2} \to 0$$

since $E[\mathbf{X}] \to \infty$ as $n \to \infty$. Secondly,

$$\frac{6\binom{n}{3}\frac{p^3+(t-1)p^4}{t}}{\left(\binom{n}{2}p^2\right)^2} \asymp \frac{n^3p^3(1/t) + n^3p^4}{n^4p^4} = \frac{1}{np\,t} + \frac{1}{n} \leq \frac{1}{np} + \frac{1}{n} \to 0$$

since $np = p/(1/n) \to \infty$ as $n \to \infty$. Finally

$$\frac{\binom{n}{2}\binom{n-2}{2}p^4 - \binom{n}{2}^2 p^4}{\left(\binom{n}{2}p^2\right)^2} \leq 0$$

.

Therefore

$$P[\mathbf{X} = 0] \to 0$$

as $n \to \infty$, for all $t \geq 1$. Therefore, whenever $p(n)/(1/n) \to \infty$ almost every graph will have an edge.

QED.

Hence, $\frac{1}{n}$ is indeed a threshold function for the appearance of edges in G.

5.1.2 $\mathbf{K_3}$: A Detailed Example

We would like to know the thresholds for the appearance of any given subgraph H in a Discrete Random Dot Product Graph. However, before we delve into the general discussion, let us consider the special case when H is K_3. First, we recall and prove Lemma 5.1.3

Lemma 5.1.3 *Let G be drawn from $\mathcal{D}[n, \{0,1\}^t, p]$. For any $u, v, w \in V(G)$ we have*

$$P[\text{the triangle } uvw] = \frac{1}{t^3}[tp^3 + 3t(t-1)p^5 + t(t-1)(t-2)p^6].$$

Proof:

Let $u, v, w \in V(G)$. Let $\mathbf{x} = [x_1, x_2, \ldots, x_t]^T$, $\mathbf{y} = y_1, y_2, \ldots, y_t]^T$, and $\mathbf{z} = [z_1, z_2, \ldots, z_t]^T$, be the vectors of u, v, and w respectively. Then

$$P[\text{the triangle } uvw] = E\left[\frac{(\mathbf{x}\cdot\mathbf{y})}{t}\frac{(\mathbf{y}\cdot\mathbf{z})}{t}\frac{(\mathbf{x}\cdot\mathbf{z})}{t}\right]$$

$$= E\left[\frac{(x_1y_1 + \cdots + x_ty_t)}{t}\frac{(y_1z_1 + \cdots + y_tz_t)}{t}\frac{(x_1z_1 + \cdots + x_tz_t)}{t}\right]$$

$$= \frac{1}{t^3}\sum_{i,j,k\in\{1,2,\ldots,t\}} E\left[x_iy_iy_jz_jx_kz_k\right]$$

and we now consider what occurs as i, j, and k vary. If $i = j = k$, then the term inside the summation is $E[x_iy_iy_jz_jx_kz_k] = E[x_i^2y_i^2z_i^2] = E[x_iy_iz_i]$, since $x_i, y_i, z_i \in \{0,1\}$. Furthermore, $E[x_iy_iz_i] = E[x_i]E[y_i]E[z_i] = p^3$ since x_i, y_i, and z_i are independent. Also, if $i = j \neq k$, then the term inside the summation is $E[x_iy_iy_jz_jx_kz_k] = E[x_iy_i^2z_ix_kz_k] = E[x_iy_iz_ix_kz_k] = E[x_i]E[y_i]E[z_i]E[x_k]E[z_k] = p^5$. In a similar manner, $E[x_iy_iy_jz_jx_kz_k] = p^5$ whenever $i = k \neq j$ and $j = k \neq i$. Finally, if i, j and k are all different, then the variables in the expected value are mutually independent and we have $E[x_iy_iy_jz_jx_kz_k] = E[x_i]E[y_i]E[y_j]E[z_j]E[x_k]E[z_k] = p^6$.

Therefore

$$P[\text{the triangle } uvw] = \frac{1}{t^3}\left(\sum_{i\in\{1,2,\ldots,t\}} p^3 + 3\sum_{\substack{i,j\in\{1,2,\ldots,t\}\\ u<v}} p^5 + \sum_{\substack{i,j,k\in\{1,2,\ldots,t\}\\ u\neq v, u\neq w, v\neq w}} p^6\right)$$

$$= \frac{1}{t^3}[tp^3 + 3t(t-1)p^5 + t(t-1)(t-2)p^6].$$

QED.

Thus, we can calculate the probability that a specific set of three vertices forms a K_3.

For any distinct $u, v, w \in V(G)$, let $Z_{u,v,w}$ be the indicator function for the triangle uvw. Then then total number of K_3's is $\mathbf{Z} = \sum_{\substack{u,v,w \in V(G) \\ u<v<w}} Z_{u,v,w}$ and we can calculate the expected number of K_3's

$$E[\mathbf{Z}] = E\left[\sum_{\substack{u,v,w \in V(G) \\ u<v<w}} Z_{u,v,w}\right] = \sum_{\substack{u,v,w \in V(G) \\ u<v<w}} E[Z_{u,v,w}]$$

$$= \binom{n}{3}\frac{1}{t^3}[tp^3 + 3t(t-1)p^5 + t(t-1)(t-2)p^6].$$

Next, we calculate the variance of $\mathbf{Z}$,

$$\text{Var}[\mathbf{Z}] = E[\mathbf{Z}^2] - E[\mathbf{Z}]^2 = E[\mathbf{Z}^2] - \binom{n}{3}^2\left(\frac{1}{t^3}\right)^2[tp^3 + 3t(t-1)p^5 + t(t-1)(t-2)p^6]^2.$$

Now let us look more closely at the first term in the variance calculation.

$$E[\mathbf{Z}^2] = E\left[\left(\sum_{\substack{u,v,w \in V(G) \\ u<v<w}} Z_{u,v,w}\right)\left(\sum_{\substack{a,b,c \in V(G) \\ a<b<c}} Z_{a,b,c}\right)\right]$$

$$= E\left[\sum_{\substack{u,v,w \in V(G) \\ u<v<w}} Z_{u,v,w}^2 + 3\sum_{\substack{u,v,w,a \in V(G) \\ u<v<w \\ a \notin \{u,v,w\}}} Z_{u,v,w}Z_{a,v,w}\right.$$

$$\left. + 3\sum_{\substack{u,v,w,a,b \in V(G) \\ u<v<w; a<b \\ \{a,b\}\cap\{u,v,w\}=\emptyset}} Z_{u,v,w}Z_{a,b,w} + \sum_{\substack{u,v,w,a,b,c \in V(G) \\ u<v<w;\, a<b<c \\ \{a,b,c\}\cap\{u,v,w\}=\emptyset}} Z_{u,v,w}Z_{a,b,c}\right]$$

$$= E\left[\sum_{\substack{u,v,w\in V(G)\\ u<v<w}} Z^2_{u,v,w}\right] + 3E\left[\sum_{\substack{u,v,w,a\in V(G)\\ u<v<w\\ a\notin\{u,v,w\}}} Z_{u,v,w}Z_{a,v,w}\right]$$

$$+3E\left[\sum_{\substack{u,v,w,a,b\in V(G)\\ u<v<w;a<b\\ \{a,b\}\cap\{u,v,w\}=\emptyset}} Z_{u,v,w}Z_{a,b,w}\right] + E\left[\sum_{\substack{u,v,w,a,b,c\in V(G)\\ u<v<w;\, a<b<c\\ \{a,b,c\}\cap\{u,v,w\}=\emptyset}} Z_{u,v,w}Z_{a,b,c}\right].$$

So, similarly to the proof of Lemma 5.1.5 we see that

$$E[\mathbf{Z}^2] = \binom{n}{3}P[\text{the triangle } uvw]$$

$$+\binom{n}{3}(n-3)3P[\{\text{the triangle } uvw\}\cap\{\text{the triangle } avw\}]$$

$$+\binom{n}{3}\binom{n-3}{2}3P[\{\text{the triangle } uvw\}\cap\{\text{the triangle } abw\}]$$

$$+\binom{n}{3}\binom{n-3}{3}P[\{\text{the triangle } uvw\}\cap\{\text{the triangle } abc\}].$$

Which, as in the proofs of Lemmas 5.1.2 and 5.1.3, can be calculated exactly and we have that

$$\mathrm{Var}[\mathbf{Z}^2] = -(1/(12t^5))((-2+n)(-1+n)n(-1+p)p^3(2t^3+2pt(-9+3n+t^2)$$

$$+2p^3(-1+t)t(-62+21n+t+t^2+t^3)+p^2(36+3n^2-18t-4t^3+6t^4+3n(-7+2t))$$

$$+2p^5(-1+t)t(346+6n^2-179t+t^2+t^3+15n(-9+4t))+2p^8(-2+t)^2(-1+t)(-108+80t-8t^2$$

$+n^2(-9+6t)+3n(21-15t+t^2))+2p^4(-1+t)(180+15n^2-80t-17t^2+t^3+t^4+3n(-35+9t+2t^2))$

$+2p^7((2-3t+t^2))(n^2(-48+33t)+3n(112-81t+3t^2+t^3)-4(144-107t+6t^2+2t^3))$

$+p^6(-1+t)(51n^2(-3+2t)+3n(357-212t-36t^2+16t^3)+2(-918+494t+161t^2-71t^3+t^4)))).$

We would like to use this variance calculation to determine a threshold function for the appearance of triangles. However, we must first set some conditions on p and t.

Let us assume that p and t have the forms

- $p = \frac{1}{n^\alpha}$
- $t = n^\beta$

where $\alpha, \beta \geq 0$. Then we propose a threshold function for the appearance of triangles in terms of α and β.

Proposition 5.1.7. *Let G be drawn from $\mathcal{D}[n, \{0,1\}^t, p]$. Let us assume that p and t have the forms $p = \frac{1}{n^\alpha}$ and $t = n^\beta$. A threshold for the appearance of triangles in G is $1/n^\alpha$ where*

$$\alpha(\beta) = \begin{cases} \frac{3-2\beta}{3} & \beta \leq \frac{3}{4} \\ \\ \frac{1}{2} & \beta \geq \frac{3}{4} \end{cases}.$$

That is, with high probability G will be triangle-free whenever $p \ll \frac{1}{n^\alpha}$ and with high probability G will have at least one triangle whenever $p \gg \frac{1}{n^\alpha}$, where α is dependent *on β and hence on t.*

Proposition 5.1.7 can be proved directly using the previous discussion, however the task is algebraically tedious and so instead we will show that Proposition 5.1.7 is a consequence of Theorem 5.3.5.

5.2 The Probability Order Polynomial

5.2.1 Defining the POP

In the previous sections, we discussed the probabilities of the appearance of an edge, a triangle and a P_3 in a Discrete Random Dot Product Graph. More generally, we would like to study the appearance of any subgraph.

Let's say that we have a graph H with m edges $E(H) = \{e_1, e_2, \ldots, e_m\}$. Let us assume that each edge $e_i \in E(H)$ has endpoints $u(e_i)$ and $v(e_i)$, so that $V(H) = \{u(e_1), v(e_1), u(e_2), v(e_2), \ldots, u(e_m), v(e_m)\}$. We wish to know the probability of H appearing as a subgraph of the random graph G (although not necessarily induced). To this end, we define, $P_{\geq}[H] = P[H$ is a subgraph of $G]$ in a specific order [1] when G is drawn from $D[n, \{0,1\}^t, p]$. Then we know by definition that the probability that H appears on a specific set of vertices is

$$P_{\geq}[H] = E\left[\left(\frac{\mathbf{x}(e_1)\cdot\mathbf{y}(e_1)}{t}\right)\left(\frac{\mathbf{x}(e_2)\cdot\mathbf{y}(e_2)}{t}\right)\cdots\left(\frac{\mathbf{x}(e_m)\cdot\mathbf{y}(e_m)}{t}\right)\right]$$

[1] Here we mean that H appears a specific set of vertices of G with the specific edges $E(H)$.

where $\mathbf{x}(e_i)$ and $\mathbf{y}(e_i)$ are the vectors of the endpoints of edge e_i, $u(e_i)$ and $v(e_i)$ respectively. So, we see that

$$P_{\geq}[H] = E\left[\left(\frac{x(e_1)_1 y(e_1)_1 + \cdots + x(e_1)_t y(e_1)_t}{t}\right)\cdots\right.$$

$$\left.\cdots\left(\frac{x(e_m)_1 y(e_m)_1 + \cdots + x(e_m)_t y(e_m)_t}{t}\right)\right]$$

$$= \frac{1}{t^m} \sum_{i_1, i_2, \ldots, i_m \in \{1,2,\ldots t\}} E[x_{i_1}(e_1) y_{i_1}(e_1)\, x_{i_2}(e_2) y_{i_2}(e_2) \cdots x_{i_m}(e_m) y_{i_m}(e_m)]$$

and we can calculate the probability by calculating the expected values of the product

$$x_{i_1}(e_1) y_{i_1}(e_1)\, x_{i_2}(e_2) y_{i_2}(e_2) \cdots x_{i_m}(e_m) y_{i_m}(e_m)$$

for all possible values of the indices $i_1, \ldots, i_m$. However, as seen in Lemmas 5.1.3 and 5.1.2, this can be a complicated task for even small graphs. Also, often we do not need to know the exact probabilities, but instead are concerned with the behavior in the limit as $n \to \infty$. So, we need to simplify the calculation of $P_{\geq}[H]$ while retaining the basic asymptotic information. To this end, we introduce the following definitions:

Definition 5.2.1. *Let $f(x, y)$ and $g(x, y)$ be real valued functions on $\mathbb{R}^2$. Then we say that*

$$f(x, y) \asymp g(x, y)$$

if $\exists c_1 > 0, \exists c_2 > 0$, so that for all x, y sufficiently small with $f(x, y) \neq 0, g(x, y) \neq 0$, *we have that* $0 < c_1 \leq \frac{f(x,y)}{g(x,y)} \leq c_2 < \infty$.

Proposition 5.2.2. *Let* $\asymp$ *be defined on real valued functions on* $\mathbb{R}^2$ *as in Definition 5.2.1. Then* $\asymp$ *is an equivalence relation. Additionally, if we have functions* f_1, f_2, g_1, *and* g_2, *positive valued functions, such that* $f_1 \asymp g_1$ *and* $f_2 \asymp g_2$ *then* $f_1 + f_2 \asymp g_1 + g_2$.

Proof:

First we show that $\asymp$ is an equivalence relation. For any real valued function, f on $\mathbb{R}^2$, $0 < 1 = \frac{f}{f}$ whenever $f \neq 0$, and so $\asymp$ is reflexive. Let f_1, f_2 be real valued function on $\mathbb{R}^2$ with $f_1 \asymp f_2$ then there exist c_1, c_2 such that $0 < c_1 \leq \frac{f_1}{f_2} \leq c_2 < \infty$. Let $a_1 = \frac{1}{c_1}$ and $a_2 = \frac{1}{c_2}$, then $0 < a_1, a_2 < \infty$ and $0 < a_2 \leq \frac{f_2}{f_1} \leq a_1$ and hence $\asymp$ is symmetric.

Finally, let f_1, f_2 and f_3 be real valued function on $\mathbb{R}^2$ with $f_1 \asymp f_2$ and $f_2 \asymp f_3$. Then there exist c_1, c_2, d_1, d_2 such that $0 < c_1 \leq \frac{f_1}{f_2} \leq c_2 < \infty$ and $0 < d_1 \leq \frac{f_2}{f_3} \leq d_2 < \infty$. Then

$$0 < c_1 d_1 \leq \frac{f_1}{f_2}\frac{f_2}{f_3} = \frac{f_1}{f_3} \leq c_2 d_2 < \infty$$

and $\asymp$ is transitive. Therefore, $\asymp$ is an equivalence relation on the set of real valued functions on $\mathbb{R}^2$.

Now, let f_1, f_2, g_1, and g_2 be positive valued functions such that $f_1 \asymp g_1$ and $f_2 \asymp g_2$. Then there exist c_1, c_2, d_1, d_2 such that $0 < c_1 \leq \frac{f_1}{g_2} \leq c_2 < \infty$ and

$0 < d_1 \leq \frac{f_2}{g_2} \leq d_2 < \infty$. First note that

$$\frac{g_1 + g_2}{f_1 + f_2} = \frac{g_1}{f_1 + f_2} + \frac{g_2}{f_1 + f_2} < \frac{g_1}{f_1} + \frac{g_2}{f_2} \leq \frac{1}{c_1} + \frac{1}{d_1} = \frac{c_1 + d_1}{c_1 d_1}.$$

Therefore taking reciprocals we have that

$$\frac{f_1 + f_2}{g_1 + g_2} \geq \frac{c_1 d_1}{c_1 + d_1} > 0.$$

Also

$$\frac{f_1 + f_2}{g_1 + g_2} = \frac{f_1}{g_1 + g_2} + \frac{f_2}{g_1 + g_2} < \frac{f_1}{g_1} + \frac{f_2}{g_2} \leq c_2 + d_2 < \infty.$$

Therefore $0 < \frac{c_1 d_1}{c_1 + d_1} \leq \frac{f_1 + f_2}{g_1 + g_2} \leq c_2 + d_2 < \infty$ and $f_1 + f_2 \asymp g_1 + g_2$.

QED.

Definition 5.2.3. *Let H be a fixed graph with h vertices and m edges. Then the Probability Order Polynomial of H, $g_H(p, 1/t)$, is a polynomial in p and $1/t$ of the form*

$$g_H(p, 1/t) = p^{x_0} + \frac{p^{x_1}}{t} + \frac{p^{x_2}}{t^2} + \cdots + \frac{p^{x_{m-2}}}{t^{m-2}} + \frac{p^{x_{m-1}}}{t^{m-1}}$$

(where $x_0, x_1, \ldots, x_{m-1} \in \mathbb{Z}_{\geq 0}$) and such that $g_H(p, 1/t) \asymp P_{\geq}[H]$.

One should note that the POP is not the only polynomial that is $\asymp$ to $P_{\geq}[H]$, however it is the only one of the form above with all the coefficients equal to one.

Let us consider for a moment how the POP of a graph H relates to the true probability of the graph appearing, $P_{\geq}[H]$. Recall that

$$P_{\geq}[H] = \frac{1}{t^m} \sum_{i_1, i_2, \ldots, i_m \in \{1,2,\ldots,t\}} E[x_{i_1}(e_1) y_{i_1}(e_1)\, x_{i_2}(e_2) y_{i_2}(e_2) \cdots x_{i_m}(e_m) y_{i_m}(e_m)]. \quad (5.1)$$

Now, consider the expectation $E[x_{i_1}(e_1) y_{i_1}(e_1)\, x_{i_2}(e_2) y_{i_2}(e_2) \cdots x_{i_m}(e_m) y_{i_m}(e_m)]$. First note that for all $1 \leq r \leq m$ the random variables $x_{i_r}(e_r)$ and $y_{i_r}(e_r)$ are independent since they are values from vectors that correspond to different vertices. Also, whenever any two of these random variables have different indices ($i_l \neq i_r$) then they are again independent since the random variables are the lth and rth coordinate of the corresponding vectors and therefore even if they correspond to the same vertex, they are still different independent random variables. In fact, the only time that two random variables within the expectation are dependent is when they refer to the same vertex and vector coordinate, for example when $x_{i_l}(e_l) = y_{i_r}(e_r)$. Whenever this occurs we can simplify the product inside the expectation by replacing $y_{i_r}(e_r)$ by $x_{i_l}(e_l)$ and then noting that $x_{i_l}(e_l) x_{i_l}(e_l) = x_{i_l}(e_l)^2 = x_{i_l}(e_l)$ since $x_{i_l}(e_l) \in \{0, 1\}$.

Let us define the following notation for each term of equation (5.1) :

- $\overrightarrow{i} = (i_1, i_2, \ldots, i_m) \in \{1, 2, \ldots, t\}^m$ is the vector of the set of indices.
- $\mathcal{L}(\overrightarrow{i}) = [x_{i_1}, y_{i_1}, x_{i_2}, y_{i_2}, \ldots, x_{i_m}, y_{i_m}]$ is the list of vector coordinates.
- $X(\overrightarrow{i})$ is the number of distinct elements in $\mathcal{L}(\overrightarrow{i})$.
- $VC(\overrightarrow{i})$ is the set of elements of $\mathcal{L}(\overrightarrow{i})$, so $X(\overrightarrow{i}) = |VC(\overrightarrow{i})|$.

Then, $E[x_{i_1}(e_1)y_{i_1}(e_1)\, x_{i_2}(e_2)y_{i_2}(e_2)\cdots x_{i_m}(e_m)y_{i_m}(e_m)] = p^{X(\overrightarrow{i})}$ in each term of Equation (5.1).

Now let $\mathcal{S} = \{i_1, i_2, \ldots, i_m\}$ the set of values of the indices in a term of Equation (5.1). Since the summation is over all possible $i_1, i_2, \ldots, i_m \in \{1, 2, \ldots, t\}$, the size of the set $\mathcal{S}$ will vary from 1 to m depending on how many distinct values are assigned to the indices. So, the probability of the graph H appearing as a subgraph of G is

$$P_{\geq}[H] = \frac{1}{t^m}\sum_{k=1}^{m}\sum_{\substack{i_1,i_2,\ldots,i_m\in\{1,2,\ldots,t\}\\ |\mathcal{S}|=|\{i_1,i_2,\ldots,i_m\}|=k}} E[x_{i_1}(e_1)y_{i_1}(e_1)\, x_{i_2}(e_2)y_{i_2}(e_2)\cdots x_{i_m}(e_m)y_{i_m}(e_m)]$$

$$= \frac{1}{t^m}\sum_{k=1}^{m}\sum_{\substack{\overrightarrow{i}\in\{1,2,\ldots,t\}^m\\ |\mathcal{S}|=|\{i_1,i_2,\ldots,i_m\}|=k}} p^{X(\overrightarrow{i})}.$$

Let $1 \leq k \leq m$ and consider an assignment of indices in which $|\mathcal{S}| = k$, say $\mathcal{S} = \{i_i, i_2, \ldots, i_m\} = \{j_1, j_2, \ldots, j_k\} \subset \{1, 2, \ldots, t\}$, where $j_1, j_2, \ldots, j_k$ are distinct . This assignment corresponds to a partition of the set of edges into k parts, $E_1, E_2, \ldots, E_k$, where edge e_l is in part E_r whenever index i_l is given the value j_r. Please note that for a specific partition $E_1, E_2, \ldots, E_k$ the choice of indices, $\overrightarrow{i}$, that induce this partition is not unique. In fact, there are $(t)_k$ different assignments of the indices that induce a specific partition of the edge set. To see this, consider a specific k-partition of the edge set, $E_1, E_2, \ldots, E_k$ and for all $1 \leq r \leq k$ let $S_r = \{i_a : e_a \in E_r\}$. Then $S = \bigcup_{r=1}^{k} S_r$ and these sets $S_1, S_2, \ldots, S_k$ define the partition. So, we do not need to know the exact values $j_1, j_2, \ldots, j_k$ associated with $S_1, S_2, \ldots, S_k$ in order to describe the partition. In fact, any of the $\binom{t}{k}$ assignments of the indices induce the same partition.

Proposition 5.2.4. *In the context of the above discussion consider some choice of the indices $i_1, i_2, \ldots, i_m \in \{1, 2, \ldots, t\}$ for which $|\mathcal{S}| = k$. Let $\mathcal{P}(\overrightarrow{i})$ be the partition of the edge set of H that corresponds to this set of index values. So, $\mathcal{P}$ has k parts, say $E_1, E_2, \ldots, E_k$, and for any two edges e_a and e_b, e_a and e_b are in the same part if and only if $i_a = i_b$. For all $1 \leq r \leq k$, let H_r be the graph that is edge induced by part E_r. Let $H_{\mathcal{P}(\overrightarrow{i})} = H_1 \oplus H_2 \oplus \cdots \oplus H_k$ be the disjoint union of the edge induced graphs. Then the number of distinct elements of $\mathcal{L}(\overrightarrow{i})$ is equal to the number of vertices in $H_{\mathcal{P}(\overrightarrow{i})}$, i.e.,*

$$X(\overrightarrow{i}) = |V(H_{\mathcal{P}(\overrightarrow{i})})|.$$

Proof:

First note that

$$|V(H_{\mathcal{P}(\overrightarrow{i})})| = |V(H_1 \oplus H_2 \oplus \cdots \oplus H_k)| = \sum_{r=1}^{k} |V(H_r)|.$$

Also, recall that $VC(\overrightarrow{i})$ is the set of vector coordinates in a term of equation (5.1). So $X(\overrightarrow{i})$, the number of distinct elements in $\mathcal{L}(\overrightarrow{i})$, is equal to $|VC(\overrightarrow{i})|$. Now, for each $1 \leq r \leq k$ let $VC(\overrightarrow{i}, j_r)$ be the set of vector coordinates in a term of equation (5.1) that correspond to the index j_r, i.e., $x_{i_a}, y_{i_a} \in VC(\overrightarrow{i}, j_r)$ if and only if $i_a = j_r$.

Then we have that $VC(\overrightarrow{i}) = \bigcup_{r=1}^{k} VC(\overrightarrow{i}, j_r)$ and so

$$X(\overrightarrow{i}) = |VC(\overrightarrow{i})| = |\bigcup_{r=1}^{k} VC(\overrightarrow{i}, j_r)| = \sum_{r=1}^{k} |VC(\overrightarrow{i}, j_r)|.$$

If we can show that $|V(H_r)| = |VC(\overrightarrow{i}, j_r)|$ for all $1 \leq r \leq k$ then $|V(H_{\mathcal{P}(\overrightarrow{i})})| = X(\overrightarrow{i})$ and we will be done. We show this by showing that the sets themselves are equal, i.e. $V(H_r) = VC(\overrightarrow{i}, j_r)$.

Suppose $v \in V(H_r)$. This means that v is an endpoint of at least one edge, say e_a, in E_r. So, $v = x_{i_a}$ or $v = y_{i_a}$. Without loss of generality, assume that $v = x_{i_a}$. Now, $v = x_{i_a} \in V(H_r)$ implies that $i_a = j_r$. Therefore $v = x_{i_a} \in VC(\overrightarrow{i}, j_r)$ and so $V(H_r) \subseteq VC(\overrightarrow{i}, j_r)$.

Similarly, suppose $v \in VC(\overrightarrow{i}, j_r)$. Then there exists at least one $a \in \{1, 2, \ldots, m\}$ with $x_{i_a}, y_{i_a} \in VC(\overrightarrow{i}, j_r)$, for which $v = x_{i_a}$ or $v = y_{i_a}$. Without loss of generality, assume that $v = x_{i_a}$. Then v is an endpoint of the edge e_a. Also, since $v = x_{i_a} \in VC(\overrightarrow{i}, j_r)$, we have that $i_a = j_r$ and the edge $e_a \in E_r$. Therefore the edge e_a is in the set that induces H_r and so the endpoints of e_a are in $V(H_r)$. Hence, $v = x_{i_a} \in V(H_r)$ and so $VC(\overrightarrow{i}, j_r) \subseteq V(H_r)$.

Therefore we have that $V(H_r) = VC(\overrightarrow{i}, j_r)$ and so

$$X(\overrightarrow{i}) = \sum_{r=1}^{k} |VC(\overrightarrow{i}, j_r)| = \sum_{j=1}^{k} |V(H_r)| = |V(H_{\mathcal{P}(\overrightarrow{i})})|.$$

QED.

So, if we think of the probability in terms of these graphs induced by edge partitions we see that

$$P_{\geq}[H] = \frac{1}{t^m} \sum_{k=1}^{m} \sum_{\substack{\overrightarrow{i} \in \{1,2,\ldots,t\}^m \\ |\mathcal{S}|=|\{i_1,i_2,\ldots,i_m\}|=k}} p^{|V(H_{\mathcal{P}(\overrightarrow{i})})|}.$$

Recalling that each partition of the edge set into k parts can occur with $\binom{t}{k}$ different choices of the index set we have that

$$P_{\geq}[H] = \frac{1}{t^m} \sum_{k=1}^{m} \sum_{\mathcal{P}(\overrightarrow{i}) \in \{\text{edge partitions of size } k\}} \binom{t}{k} p^{|V(H_{\mathcal{P}(\overrightarrow{i})})|}$$

$$\asymp \sum_{k=1}^{m} \sum_{\mathcal{P}(\overrightarrow{i}) \in \{\text{edge partitions of size } k\}} \frac{1}{t^{m-k}} p^{|V(H_{\mathcal{P}(\overrightarrow{i})})|}.$$

One should note that for each k there are $S(m,k)$ partitions $\mathcal{P}(\overrightarrow{i})$ of size k, where $S(m,k)$ is the Stirling number of the second kind.

Now since $p \in [0,1]$ for each k we see that

$$\sum_{\mathcal{P}(\overrightarrow{i}) \in \{\text{edge partitions of size } k\}} \frac{1}{t^{m-k}} p^{|V(H_{\mathcal{P}(\overrightarrow{i})})|} \asymp \frac{1}{t^{m-k}} p^{h_k}$$

where

$$h_k = \min_{\overrightarrow{i} : X(\overrightarrow{i})=k} |V(H_{\mathcal{P}(\overrightarrow{i})})| = \sum_{j=1}^{k} |V(H_j)|$$

is the minimum number of vertices used in a graph induced by an edge partition of size k. Therefore

$$P_{\geq}[H] \asymp \sum_{k=1}^{m} \frac{1}{t^{m-k}} p^{h_k}.$$

The above discussion gives us the following lemma.

Lemma 5.2.5. *Let H be a graph with $|E(H)| = m$. Let*

$$g_H(p, 1/t) = p^{x_0} + \frac{p^{x_1}}{t} + \frac{p^{x_2}}{t^2} + \cdots + \frac{p^{x_{m-2}}}{t^{m-2}} + \frac{p^{x_{m-1}}}{t^{m-1}}$$

be the POP of H. Then for all $0 \leq j \leq m-1$, $x_j = h_{m-j}$, the minimum number of vertices used in a graph induced by a partition of the edge set into of $m-j$ parts.

One should note that for any graph H, Lemma 5.2.5 gives us that $x_0 \geq x_1 \geq \cdots \geq x_{m-1}$.

5.2.2 POPs of Trees, Cycles, and Complete Graphs

While g_H is simpler to calculate then the true value of $P_{\geq}[H]$, it is still often complicated. So we will fist consider a few classes of graphs and the POPs that they produce. We will begin with the often simple class of trees, but first we need the following results.

Proposition 5.2.6. *Let H be a graph on n vertices and m edges without isolated vertices. Let*

$$g_H(p, 1/t) = p^{x_0} + \frac{p^{x_1}}{t} + \frac{p^{x_2}}{t^2} + \cdots + \frac{p^{x_{m-2}}}{t^{m-2}} + \frac{p^{x_{m-1}}}{t^{m-1}}$$

be the POP of H. Then $x_0 = 2m$ and $x_{m-1} = h$.

Proof:

By Lemma 5.2.5, $x_0 = h_{m-0} = h_m$ and is the minimum number of vertices used in a graph induced by a partition of the edge set into m parts. The only way to partition $E(H)$ into m parts is to place each edge into its own part. Therefore each H_j, $1 \leq j \leq m$, consists of a single edge and $x_0 = h_m = 2m$. Likewise, $x_{m-1} = h_{m-(m-1)} = h_1$ and the only way to partition $E(H)$ into 1 part is $E(H)$ itself which induces H. Therefore $x_{m-1} = h_0 = h$.

QED.

Lemma 5.2.7. *Let H be a graph on h vertices with m edges without isolated vertices. Let $\hat{H}$ be another graph formed from H by adding a new pendant edge to H. If the POP of H is*

$$g_H(p, 1/t) = p^{x_0} + \frac{p^{x_1}}{t} + \frac{p^{x_2}}{t^2} + \cdots + \frac{p^{x_{m-2}}}{t^{m-2}} + \frac{p^{x_{m-1}}}{t^{m-1}}.$$

Then, the POP of $\hat{H}$ is

$$g_{\hat{H}}(p, 1/t) = p^{x_0+2} + \sum_{i=1}^{m-1} \frac{p^{\min\{x_i+2, x_{i-1}+1\}}}{t^i} + \frac{p^{x_{m-1}+1}}{t^m}.$$

Proof:

Let H be a graph on the vertex set $V(H)$ (where $|V(H)| = h$) and having the edge set $E(H) = \{e_1, e_2, \ldots, e_m\}$. Let H have POP

$$g_H(p, 1/t) = p^{x_0} + \frac{p^{x_1}}{t} + \frac{p^{x_2}}{t^2} + \cdots + \frac{p^{x_{m-2}}}{t^{m-2}} + \frac{p^{x_{m-1}}}{t^{m-1}}.$$

Let v be some vertex $v \in V(H)$ and $\hat{H}$ be a graph on $V(\hat{H}) = V(H) \bigcup \{\hat{v}\}$, $\hat{v} \notin V(H)$, with edge set $E(\hat{H}) = E(H) \bigcup \{\{v, \hat{v}\}\}$. So $\hat{H}$ is a new graph formed from H by adding a pendant edge to the vertex $v \in V(H)$. We wish to calculate the POP of $\hat{H}$.

Now, $\hat{H}$ has exactly one more edge than H and so $|E(\hat{H})| = |E(H)| + 1 = m + 1$. Therefore by Lemma 5.2.5 we know that the POP of $\hat{H}$ has the form

$$g_{\hat{H}}(p, 1/t) = p^{y_0} + \frac{p^{y_1}}{t} + \frac{p^{y_2}}{t^2} + \cdots + \frac{p^{y_{m-1}}}{t^{m-1}} + \frac{p^{y_m}}{t^m}$$

where $y_j = h_{(m+1)-j}$ and is the smallest number of vertices in a graph induced by a partition of the edge set of $\hat{H}$ into $(m+1) - j$ parts. So to calculate $g_{\hat{H}}(p, \frac{1}{t})$, we need to partition the edges of $\hat{H}$ into partitions of size $1, 2, \ldots, m, m+1$. This can be done by first partitioning the edges $e_1, e_2, \ldots, e_m$ into partitions of size $1, 2, \ldots, m$ and then either place the edge $\{v, \hat{v}\}$ into one of the existing parts or giving $\{v, \hat{v}\}$ its own new part. Note that partitioning the edges $e_1, e_2, \ldots, e_m$ is the same as partitioning the edges of H and that by Lemma 5.2.5 we know that each x_j in the POP of H is the smallest number of vertices in a graph induced by a partition of the edge set of H into $m - j$ parts. Let us calculate $y_0, y_1, \ldots, y_m$.

Step 1: Calculate y_0.

By Proposition 5.2.6, $y_0 = 2|E(\hat{H})| = 2(m+1)$ and $x_0 = 2|E(H)| = 2m$, therefore $y_0 = x_0 + 2$.

Step 2: Calculate y_j for $1 \leq j \leq m-1$, partitioning the edge set of $\hat{H}$ into $(m+1)-j$ parts.

We can obtain all of the partitions of the edge set of $\hat{H}$ by partitioning the edges of H into $(m+1)-j-1 = m-j$ parts and then either placing the edge $\{v, \hat{v}\}$ into its own part, an existing part containing v, or an existing part that does not contain v. We will consider each of these cases separately.

First, for any $1 \leq j \leq m-1$, let us consider obtaining a partition of the edge set of $\hat{H}$ into $(m+1)-j$ parts by partitioning the edges of H into $(m+1)-i-j = m-j$ parts, $H_1, H_2, \ldots, H_{m-j}$, and then placing the edge $\{v, \hat{v}\}$ into its own part. Then any partition that is formed this way induces a graph on $\sum_{r=1}^{m-j} H_r + |\{v, \hat{v}\}| = x_j + 2$ vertices. Therefore, $y_j \leq x_j + 2$.

Now, for any $1 \leq j \leq m-1$, consider a partition of the edge set of $\hat{H}$ into $(m+1)-j$ parts obtained from a partition of the edges of H into $(m+1)-j$ parts by placing $\{v, \hat{v}\}$ into an existing part. Note that the number of vertices used in a graph induced by just the partition of the edges of H into $(m+1)-j = m-(j-1)$ parts is x_{j-1}. Now, if $\{v, \hat{v}\}$ is placed into an existing part that does not contain v then the graph that the partition induces will contain 2 additional vertices and so the total number of vertices used will be $x_{j-1}+2$ and so $y_j \leq x_{j-1}+2$.

Finally, since $v \in V(H)$ and H does not contain isolated vertices, we know that in every partition of the edges of H we can always find a part that contains v. So, if $\{v, \hat{v}\}$ is placed into an existing part that contains v, then the total number of vertices that is used in a graph induced by the partition only increases by one. Therefore in the partition that uses the minimum number of vertices we can always place $\{v, \hat{v}\}$ in a part already containing v and we have that $y_j \leq x_{j-1} + 1$.

Since all partitions of $\hat{H}$ can be obtained from partitions of H in the above manner we see that the smallest number of vertices required in a partitioning of the edge set of $\hat{H}$ into $(m+1) - j$ parts is $y_j = \min\{x_j + 2, x_{j-1} + 1\}$.

Step 3: Calculate y_m.

By Proposition 5.2.6, $y_m = |V(\hat{H})| = h + 1$ and $x_{m-1} = |V(H)| = h$, therefore $y_m = x_{m-1} + 1$.

Thus, the POP of $\hat{H}$ is

$$g_{\hat{H}}(p, 1/t) = p^{x_0+2} + \sum_{i=1}^{m-1} \frac{p^{\min\{x_i+2, x_{i-1}+1\}}}{t^i} + \frac{p^{x_{m-1}+1}}{t^m}.$$

QED.

Theorem 5.2.8. *Let T be a tree on $h \geq 2$ vertices then the POP of T is*

$$g_T(p, \frac{1}{t}) = p^{2(h-1)} + \frac{p^{2(h-1)-1}}{t} + \cdots + \frac{p^{h+1}}{t^{h-3}} + \frac{p^h}{t^{h-2}} = \sum_{i=0}^{h-2} \frac{p^{2(h-1)-i}}{t^i}.$$

Proof:

The proof is by induction on the number of vertices h. Note that by Proposition

5.1.1 and Lemma 5.1.2 we see that the theorem holds when $h = 2, 3$.

Assume that the theorem holds for all trees on $h-1$ vertices and let us consider a tree, T, on h vertices. Now, since T is a tree, it must contain at least two leaves. Let v be a leaf of T. Then the subgraph H formed by removing the leaf v from T is a tree on $h-1$ vertices. Therefore the POP of H is

$$g_H(p, \frac{1}{t}) = p^{2[(h-1)-1]} + \frac{p^{2[(h-1)-1]-1}}{t} + \cdots + \frac{p^{[(h-1)-1]+2}}{t^{[(h-1)-1]-2}} + \frac{p^{[(h-1)-1]+1}}{t^{[(h-1)-1]-1}}$$

by the induction hypothesis. Now, we use Lemma 5.2.7 to find the POP of T from the POP of H. In the context of Lemma 5.2.7, $x_i = 2[(h-1)-1] - i = 2(h-1) - (i+2)$ and $e = h-2$ and so

$$f_T(p, 1/t) = p^{x_0+2} + \sum_{i=1}^{e-1} \frac{p^{\min\{x_i+2, x_{i-1}+1\}}}{t^i} + \frac{p^{x_{e-1}+1}}{t^e}$$

$$= p^{2(h-1)-(0+2)+2} + \sum_{i=1}^{(h-2)-1} \frac{p^{\min\{2(h-1)-(i+2)+2, 2(h-1)-((i-1)+2)+1\}}}{t^i} + \frac{p^{2(h-1)-((h-2)-1+2)+1}}{t^{h-2}}$$

$$= p^{2(h-1)} + \sum_{i=1}^{h-3} \frac{p^{\min\{2(h-1)-i, 2(h-1)-i+1-2+1\}}}{t^i} + \frac{p^{(h-1)+1}}{t^{h-2}}$$

$$= p^{2(h-1)} + \sum_{i=1}^{h-3} \frac{p^{\min\{2(h-1)-i, 2(h-1)-i\}}}{t^i} + \frac{p^h}{t^{h-2}}$$

$$= p^{2(h-1)} + \sum_{i=1}^{h-3} \frac{p^{2(h-1)-i}}{t^i} + \frac{p^h}{t^{h-2}}$$

which is the desired result. Therefore, by induction, the theorem holds for all trees on two or more vertices.

QED.

Theorem 5.2.9. *Let C_h be a cycle on $h \geq 3$ vertices then the POP of C_h is*

$$g_{C_h}(p, \frac{1}{t}) = p^{2h} + \frac{p^{2h-1}}{t} + \cdots + \frac{p^{h+2}}{t^{h-2}} + \frac{p^h}{t^{h-1}} = \sum_{k=2}^{h} \frac{p^{h+k}}{t^{h-k}} + \frac{p^h}{t^{h-1}}.$$

For example the POP of C_5 is $g_{C_5}(p, \frac{1}{t}) = p^{10} + \frac{p^9}{t} + \frac{p^8}{t^2} + \frac{p^7}{t^3} + \frac{p^5}{t^4}$.

Proof:

Recall that by Definition 5.2.3 and Proposition 5.2.6 the POP of C_h has the general form

$$g_{C_h}(p, 1/t) = p^{2h} + \frac{p^{x_1}}{t} + \frac{p^{x_2}}{t^2} + \cdots + \frac{p^{x_{h-2}}}{t^{h-2}} + \frac{p^h}{t^{h-1}}.$$

We would like to show that

$$g_{C_h}(p, 1/t) = p^{2h} + \sum_{k=2}^{h-1} \frac{p^{h+k}}{t^{h-k}} + \frac{p^h}{t^{h-1}}.$$

In other words, we would like to show that for all $2 \leq k \leq h-1$, $x_{h-k} = h+k$, which by Lemma 5.2.5 is equivalent to showing that the minimum number of vertices used in a graph induced by a partition of the edge set into k parts is $h_k = h+k$.

Let us consider when the edges are partitioned into $2 \leq k \leq h-1$ parts.

Claim: The partition of the edge set into a single path $P_{h-(k-2)}$ and $(k-1)$ K_2s induces a graph that uses the minimum number of vertices h_k.

The partition of the edge set into a single $P_{h-(k-2)}$ and $(k-1)$ K_2s induces a graph that uses $h-(k-2)+(k-1)2 = h-k+2+2k-2 = h+k$ vertices, therefore if the claim is true then $h_k = h+k$ and we are done.

Proof of claim by smallest counterexample: First consider when $k = 2$. We wish to partition the edge set into 2 parts. If we partition the edge set into a P_h and a K_2 we use $h+2$ vertices. We claim that this is best. Note that for any partition of the edge set into 2 parts, say E_1 and E_2 the components of each part will be paths. Therefore, for each $i = 1, 2$, the number of vertices in a graph H_i induced by E_i will be $|E(H_i)|$+(number of components in H_i). Since each part induces a graph with at least one component, the total number of vertices used in a graph induced by this partition is at least $|E(H_1)|+1+|E(H_2)|+1 = |E(C_h)|+2 = h+2$. Hence $h_k = h+2$ and the claim is true for the base case $k = 2$.

Next, let $2 < k \leq h-1$ be the smallest value for which the claim fails. So, by our induction assumption, a partition of the edges of C_h into $k-1$ parts that induces a graph with the minimum number of vertices is the partition of C_h into a $P_{h-((k-1)-2)}$ path and $(k-2)K_2$s which uses $h_{k-1} = h-k-3+2(k-2) = h+k-1$ vertices.

Also, by our assumption the partition of the edges set into a P_{h-k-2} and $(k-1)K_2$s which uses $h+k$ vertices is not the best case. Therefore, there exists a k-partition of the edges of C_n that induces a graph that uses at most $h+k-1$ vertices. Now consider one of these best k-partitions, say $E_1, E_2, \ldots, E_k$. We know that each part of this partition induces a graph whose components are paths. Therefore there must be a leaf u in H_1. Now, u has degree two in C_h, therefore u must appear in some $H_i\,(i \neq 1)$. Combine E_1 and E_i into a single part and create a new partition that contains this union and the remaining parts $E_j, (j \neq 1, i)$. This new partition induces a graph that uses at least 1 fewer vertices than the original partition since u appears only once. Therefore the new partition uses at most $h+k-1-1 = h+k-2$ vertices, however it has $k-1$ parts. This is a contradiction since we have already determined that the best $k-1$ partitioning induces a graph that uses $h+k-1$ vertices. Therefore our assumption is false and no such smallest k exists. Hence for all $2 \leq k \leq h-1$ the claim holds.

Therefore the POP of C_h is indeed $g_{C_h}(p, 1/t) = p^{2h} + \sum_{k=2}^{h-1} \frac{p^{h+k}}{t^{h-k}} + \frac{p^h}{t^{h-1}}$.

QED.

Now let us move onto a harder class of graphs: the set of complete graphs. We prove the following lemma before we state the main result.

Lemma 5.2.10. *Let H be a connected graph and for all $1 \leq k \leq |E(H)|$, let h_k be the minimum number of vertices used in a graph induced by a partition of the edge set of H into k parts. Then $h_{k-1} < h_k$.*

Proof:

Let H be a connected graph. Let $\mathcal{P}_k$ be a partition of the edges of H into k parts that induces a graph with h_k vertices. Since H is connected, there exists parts E_i and E_j that induce graphs H_i and H_j for which $V(H_i) \bigcap V(H_j) \neq \emptyset$. Therefore, there exists $v \in V(H_i) \bigcap V(H_j)$. We can obtain a partition of the edges of H into $k-1$

parts by taking the union $E_i \bigcup E_j$ as a single part and combining it with the other $k-2$ parts from our original k-partition. This new partition will induce a graph with at least one fewer vertex since v now appears only once in $E_i \bigcup E_j$. Therefore $h_{k-1} \leq h_k - 1$.

QED.

Theorem 5.2.11. *Let K_h be the complete graph on $h \geq 4$ vertices and let $1 \leq k \leq \binom{h}{2}$. Then a partition of the edge set, $E(K_h)$, into k parts that induces a graph with the minimum number of vertices is found by partitioning $E(K_h)$ into $k-1$ single edge sets, $E_1, E_2, \ldots, E_{k-1}$, and one large set, E_k, of size $\binom{h}{2} - (k-1)$ that induces a graph H_k that contains a clique of size c_k (where c_k maximizes $\binom{c_k}{2} \leq \binom{h}{2} - (k-1)$) and at most one other vertex incident to $\binom{h}{2} - (k-1) - \binom{c_k}{2}$ edges. Therefore, $|V(H_k)| = b_k$ where $\binom{b_k - 1}{2} < \binom{h}{2} - (k-1) \leq \binom{b_k}{2}$ and the POP of K_h is*

$$g_{K_h}(p, 1/t) = \sum_{k=1}^{\binom{h}{2}} \frac{p^{b_k + 2(k-1)}}{t^{\binom{h}{2} - k}}.$$

Proof:

By Definition 5.2.3 of the POP and Lemma 5.2.5, we know that $g_{K_h}(p, 1/t)$ has the form

$$g_{K_h}(p, 1/t) = \sum_{i=0}^{\binom{h}{2}-1} \frac{p^{x_i}}{t^i}$$

where $x_i = h_{\binom{h}{2}-i}$ is the minimum number of vertices used in a graph induced by a partition of the edges set into $\binom{h}{2} - i$ parts. We would like to show that

$$g_{K_h}(p, 1/t) = \sum_{i=0}^{\binom{h}{2}-1} \frac{p^{h_{\binom{h}{2}-i}}}{t^i} = \sum_{k=1}^{\binom{h}{2}} \frac{p^{b_k+2(k-1)}}{t^{\binom{h}{2}-k}}. \tag{5.2}$$

In the second summation of Equation 5.2 we have reversed the order of the terms. We can reorder the first summation by using the change of index $k = \binom{h}{2} - i$. Then

$$g_{K_h}(p, 1/t) = \sum_{k=1}^{\binom{h}{2}} \frac{p^{x_{\binom{h}{2}-k}}}{t^{\binom{h}{2}-k}} = \sum_{k=1}^{\binom{h}{2}} \frac{p^{b_k+2(k-1)}}{t^{\binom{h}{2}-k}}.$$

Showing that Equation 5.2 holds is equivalent to showing that for all $1 \leq k \leq \binom{h}{2}$, $b_k + 2(k-1) = x_{\binom{h}{2}-k} = h_k$ and is the minimum number of vertices used in a graph induced by a partition of the edge set into k parts.

The partition as described in the statement of the theorem induces a graph that uses $b_k + 2(k-1)$ vertices, since $|V(H_1)| = |V(H_2)| = \cdots = |V(H_{k-1})| = 2$ and $|V(H_k)| = b_k$. So, if we can show that this partition does indeed minimize the number of vertices used in an induced graph, then $h_k = b_k + 2(k-1)$ and the theorem holds. The rest of the proof is devoted to showing this fact.

First note that such a partition exists and is easily obtained from K_h by systematically removing edges incident with v_1 until no more remain and then repeating with the edges incident to v_2 and so forth, until $k-1$ edges have been removed.

We will now prove this theorem by the smallest counter example technique. Let us first consider the base cases when $k = 1, 2$.

When $k = 1$, we partition the edge set into a single part $E_1 = E(K_h)$. Therefore the graph induced by this partition is K_h and has h vertices. Hence, the minimum number of vertices used in a partition of the edge set into 1 part is $h_1 = h$. Also, b_1 is such that $\binom{b_1-1}{2} < \binom{h}{2} - (1-1) = \binom{h}{2} \leq \binom{b_1}{2}$. Therefore, $b_1 = h$ and so $b_k + 2(k-1) = b_1 + 2(1-1) = b_1 = h = h_1$ and the theorem holds when $k = 1$.

Let $k = 2$. Then we need $\binom{b_2-1}{2} < \binom{h}{2} - (2-1) \leq \binom{b_2}{2}$, therefore recalling that $h \geq 4$, it follows that $b_2 = h$ and we want to show that $h_2 = h + 2$. Let us assume otherwise, that is let us assume that there exists a partition of the edge set into 2 parts that uses at most $h+1$ vertices. Each part must induce a graph that contains at least 3 vertices, otherwise the partition would be of the type described in the statement of the theorem and would use $h+2$ vertices. Let parts E_1 and E_2 induce graphs H_1 and H_2, respectively. Also, let $r_i = |V(H_i)|$, $i = 1, 2$, then $r_1 + r_2 \leq h + 1$ and $r_1, r_2 \geq 3$.

Now, consider a vertex $u \in V(K_h)$, u must by in one or both of H_1 and H_2. Let us first consider the case when u is in only one of the induced graphs. Without loss of generality assume that $u \in H_1$ and $u \notin H_2$. Now, since the graph is complete, u is adjacent to all of the other $h-1$ vertices in $V(K_h) - \{u\} = \{v_1, v_2, \ldots v_{k-1}\}$ in K_h. Therefore the edges $uv_1, uv_2, \ldots, uv_{k-1}$ must appear in the partition, and since $u \notin H_2$, these edges must all be in E_1. Therefore, $v_1, v_2, \ldots, v_{k-1} \in V(H_1)$ and so $r_1 = h$. Hence $r_1 + r_2 \geq h + 3$ which is a contradiction. Therefore u must appear in both H_1 and H_2.

Let us assume that $u \in V(H_1)$ and $u \in V(H_2)$. Now, $u \in V(H_2)$ implies that there is some other vertex $v \in H_2$ with $u \sim v$ and $uv \in E_2$. Now, if v is only in $V(H_2)$ and not in $V(H_1)$ then by the same argument as above, all of the vertices must be in $V(H_2)$ and so $r_2 = h$, implying again that $r_1 + r_2 \geq h+3$ and we have a contradiction. Therefore v must also appear in both induced graphs. Now all h vertices in K_h must appear in at least one of the induced graphs and u and v appear in both of them, therefore $r_1 + r_2 \geq h + 2$ which again is a contradiction.

So, there cannot exist a partition of the edge set that induces a graph with fewer than $h + 2 = b_2 + 2$ vertices and so our theorem holds when $k = 2$.

Let us consider the smallest k for which the theorem does not hold (noting that $k \geq 3$). So, there exists a partition $E_1, E_2, \ldots, E_k$ of the edge set into k parts that induces a graph on $h_k < b_k + 2(k-1)$ vertices, however, the theorem holds for all partitions of size less than k and so $h_{k-1} = b_{k-1} + 2(k-2)$. Now by Lemma 5.2.10, we have that $h_k \geq h_{k-1} + 1 = b_{k-1} + 2(k-2) + 1$. Also, by definition of b_k it is easy to see that $b_k \leq b_{k-1} \leq b_k + 1$ and so we have that

$$b_k + 2(k-1) > h_k \geq h_{k-1} + 1 = b_{k-1} + 2(k-2) + 1 \geq b_k + 2(k-1) - 1.$$

Therefore $h_k = h_{k-1} + 1$ and we have the following claims.

Claim 1: $|V(H_i) \bigcap V(H_j)| \leq 1$ for all $i \neq j$.

Proof of Claim 1: Suppose that there exists parts E_i and E_j which induce graphs H_i and H_j for which $|V(H_i) \bigcap V(H_j)| \geq 2$. Then we can obtain a partition of size $k-1$ from this partition of size k by taking the union $E_i \bigcup E_j$ as a single part and combining it with the other $k-2$ parts from our k-partition. This new partition will use at most $h_k - 2$ vertices since there are two vertices in the intersection of $V(H_i)$ and $V(H_j)$ that appeared twice in the k-partition and will only appear once in the new partition. Therefore $h_{k-1} \leq h_k - 2 = h_{k-1} + 1 - 2 = h_{k-1} - 1$ and we have a contradiction. Hence, Claim 1 holds.

Claim 2: Each induced graph H_i a clique.

Proof of Claim 2: Suppose that there exists a part E_i that induces a graph H_i that is not a clique. Therefore, there exist $u, v \in V(H_i)$ for which $uv \notin E_i$. Now, since the graph is complete, the edge uv must be in some part E_j, $i \neq j$. So, $u, v \in V(H_j)$ and $V(H_i) \bigcap V(H_j) \geq 2$ which contradicts Claim 1. Hence, Claim 2 holds.

Now, without loss of generality, let us assume that the the partition is such that $r_1 \geq r_2 \geq \cdots \geq r_k$. Let us consider H_1 and H_2. Note that $r_1 \geq r_2 \geq 3$ otherwise if $2 \geq r_2 \geq \cdots \geq r_k$ we would have a partition of the type described in the statement of the theorem and $h_k = b_k + 2(k-1)$. Let us assume that $V(H_1) \bigcap V(H_2) = \emptyset$. Then there exists a vertex $u \in V(H_2)$, with $u \notin V(H_1)$ and consider the new partition $\hat{E}_1, \hat{E}_2, \ldots, \hat{E}_k$ formed from $E_1, E_2, \ldots, E_k$ in the following way:

Step 1. Let $\hat{E}_2 = E_2 - \{uv : v \neq u, uv \in E_2\}$, i.e., $\hat{E}_2$ is formed from E_2 by deleting all edges incident with u. This step removes u from H_2 and $V(\hat{H}_2) = V(H_2) - \{u\}$. Note that since $r_2 \geq 3$ there is at least 2 edges in $\hat{E}_2$.

Step 2. Let $V(H_1) = \{x_1, x_2, \ldots, x_{r_1}\}$ and consider the parts $E_{i_1}, E_{i_2}, \ldots, E_{i_{r_2 - 1}}$

that contain the edges $ux_1, ux_2, \ldots, ux_{r_2-1}$ respectively. Note that $E_1, E_2 \notin \{E_{i_1}, E_{i_2}, \ldots, E_{i_{r_2-1}}\}$ since $u \notin V(H_1)$ and none of the $x_i's$ are in $V(H_2)$. Also note that whenever $a \neq b$, $E_{i_a} \neq E_{i_b}$, otherwise $ux_a, ux_b \in E_{i_a}$ and $|V(H_1) \bigcap V(H_{i_a})| = 2$ which contradicts Claim 1. We now let $\hat{E}_1 = E_1 \bigcup E_{i_1} \bigcup \cdots E_{i_{r_2-1}}$ so that $\hat{H}_1$ contains $H_1, H_{i_1}, \ldots, H_{i_{r_2-1}}$ as subgraphs.

Step 3. Place the $r_2 - 1$ edges in $\{uv : v \neq u, uv \in E_2\}$ into single edge sets $\hat{E}_{i_1}, \hat{E}_{i_2}, \ldots, \hat{E}_{i_{r_2-1}}$.

Step 4. For all $j \notin \{1, 2, i_1, i_2, \ldots, i_{r_2-1}\}$ let $\hat{E}_j = E_j$.

This new partition has exactly k parts, $\hat{E}_1, \hat{E}_2, \ldots, \hat{E}_k$. We want to show that it also induces a graph on h_k vertices.

First note that $V(\hat{H}_2) = V(H_2) - \{u\}$ and so $\hat{r}_2 = r_2 - 1$. Next, $\hat{r}_{i_1} = \hat{r}_{i_2} = \cdots = \hat{r}_{i_{r_2-1}} = 2$ since the graphs are induced by a single edge. Also note that for all $j \notin \{1, 2, i_1, i_2, \ldots, i_{r_2-1}\}$ let $\hat{r}_j = r_j$ since the graphs are the same. Finally we have the following claim:

Claim 3: $\hat{r}_1 = r_1 + 1 + \sum_{l=1}^{r_2-1} r_{i_l} - 2(r_2 - 1)$.

Proof of Claim 3: $V(\hat{H}_1) = V(H_1) \bigcup \{u\} \bigcup_{l=1}^{r_2-1} V(H_{i_l})$ by definition. Also $|V(H_1) \bigcup \{u\}| = r_1 + 1$ since $u \notin V(H_1)$. Now for all l, $V(H_1) \bigcap V(H_{i_l}) = \{x_l\}$ since $x_l \in V(H_1)$ and $x_l \in V(H_{i_l})$ and by Claim 1 there intersection contains at most 1 vertex. Also, for all $a \neq b$, $V(H_{i_a}) \bigcap V(H_{i_b}) = \{u\}$ since $u \in V(H_{i_a})$ and $u \in V(H_{i_b})$ and by Claim 1 there intersection contains at most 1 vertex. So each vertex in $V(\hat{H}_1) = V(H_1) \bigcup \{u\} \bigcup_{l=1}^{r_2-1} V(H_{i_l})$ appears exactly once in the sets

$V(H_1), \{u\}, V(H_{i_1}), V(H_{i_2}), \ldots, V(H_{i_{r_2-1}})$ except for u and $x_1, x_2, \ldots, x_{r_2-1}$. Now, u appears in each of $V(H_{i_l})$ as well as as in $\{u\}$ and so u appears exactly $1 + r_2 - 1$ times. Each of the $x_1, x_2, \ldots, x_{r_2-1}$ appear exactly twice, once in $V(H_1)$ and once in their respective $V(H_{i_l})$'s. Therefore when we sum the sizes of the sets we over count by $r_2 - 1$ for u and $r_2 - 1$ for each of the x_l's. Hence $\hat{r}_1 - r_1 + 1 + \sum_{l=1}^{r_2-1} r_{i_l} - 2(r_2 - 1)$ and the claim in proven.

So the number of vertices used in a graph induced by the new partition is

$$\sum_{j=1}^{k} \hat{r}_j = \left[r_1 + 1 + \sum_{l=1}^{r_2-1} r_{i_l} - 2(r_2 - 1) \right] + [r_2 - 1] + [2(r_2 - 1)] + \left[\sum_{l \notin \{1,2,i_1,i_2,\ldots,i_{r_2-1}\}} r_l \right]$$

$$= \sum_{j=1}^{k} r_j + 1 - 2(r_2 - 1) - 1 + 2(r_2 - 1)$$

$$= \sum_{j=1}^{k} r_j$$

$$= h_k.$$

Therefore this partition also induces a graph on the minimum number of vertices. However, $\hat{H}_1$ is not a clique since only $r_2 - 1 \leq r_1 - 1 < r_1$ edges exist between u and other vertices in $V(\hat{H}_1)$ and there are at least the original r_1 other vertices from $V(H_1)$ in $V(\hat{H}_1)$. This fact contradicts Claim 2. Therefore h_k cannot be less than $b_k + 2(k - 1)$ and the theorem holds if the intersection of $V(H_1)$ and $V(H_2)$ is empty.

Let us assume that $V(H_1) \bigcap V(H_2) \neq \emptyset$. Therefore by Claim 1 they intersect at exactly one vertex w. Here as above, there exists $u \in V(H_2)$ with $u \notin V(H_1)$. We will create the new k-partition in the same way as above, except that we will put the edge uw into $\hat{E}_1$. So, in Step 3, we will only have $r_2 - 2$ single edge sets (note $r_2 - 2 \leq 1$ since $r_2 \leq 3$). And in Step 2 we assume that $V(H_1) = \{w, x_1, \ldots, x_{h-1}\}$ and we add the edge uw to $\hat{E}_1$ and only need to add $r_2 - 2$ additional parts $E_{i_1}, E_{i_2}, \ldots, E_{i_{r_2-1}}$. Then Claim 3 becomes:

Claim 3′: $\hat{r}_1 = r_1 + 1 + \sum_{l=1}^{r_2-2} r_{i_l} - 2(r_2 - 2)$

The proof of Claim 3′ is the similar to that of Claim 3 and the number of vertices used in a graph induced by the new partition is

$$\sum_{j=1}^{k} \hat{r}_j = \left[r_1 + 1 + \sum_{l=1}^{r_2-2} r_{i_l} - 2(r_2 - 2) \right] + [r_2 - 1] + [2(r_2 - 2)] + \left[\sum_{l \notin \{1,2,i_1,i_2,\ldots,i_{r_2-2}\}} r_l \right]$$

$$= \sum_{j=1}^{k} r_j + 1 - 2(r_2 - 2) - 1 + 2(r_2 - 2)$$

$$= \sum_{j=1}^{k} r_j$$

$$= h_k.$$

Again this new partition induces a graph on the minimum number of vertices. However, $\hat{H}_1$ is not a clique since only $r_2 - 2 + 1 \leq r_1 - 1 < r_1$ edges exist between u and other vertices in $V(\hat{H}_1)$ and there are at least the original r_1 other vertices from $V(H_1)$ in $V(\hat{H}_1)$. This fact contradicts Claim 2. Therefore h_k cannot be less than $b_k + 2(k-1)$ and the theorem holds if the intersection of $V(H_1)$ and $V(H_2)$is nonempty.

Hence there is no smallest k for which there exists a partition that induces a graph on less than $b_k + 2(k-1)$ vertices. Therefore $h_k = b_k + 2(k-1)$ and the Theorem holds.

QED.

So, for example, the POP of K_4 is $g_{K_4}(p, 1/t) = p^{12} + \frac{p^{11}}{t} + \frac{p^9}{t^2} + \frac{p^8}{t^3} + \frac{p^6}{t^4} + \frac{p^4}{t^5}$.

Corollary 5.2.12. *Let K_h be the complete graph on $h \geq 2$ vertices. Then the POP if K_h is*

$$g_{K_h}(p, 1/t) = \sum_{k=1}^{\binom{h}{2}} \frac{p^{b_k+2(k-1)}}{t^{\binom{h}{2}-k}}$$

where b_k satisfies $\binom{b_k-1}{2} < \binom{h}{2} - (k-1) \leq \binom{b_k}{2}$.

Proof:

Let $h = 2$ so that K_2 is a single edge. Then by Proposition 5.1.1 we know that $g_{K_2}(p, 1/t) = p^2$ and we must show that $p^2 = \sum_{k=1}^{\binom{h}{2}} \frac{p^{b_k+2(k-1)}}{t^{\binom{h}{2}-k}}$. First $\binom{h}{2} = \binom{2}{2} = 1$ and so $\sum_{k=1}^{\binom{h}{2}} \frac{p^{b_k+2(k-1)}}{t^{\binom{h}{2}-k}} = p^{b_1}$. Also, b_1 must satisfy $\binom{b_1-1}{2} < \binom{h}{2} - (1-1) = 1 \leq \binom{b_1}{2}$, and so $b_1 = 2$. Therefore $g_{K_2}(p, 1/t) = \sum_{k=1}^{\binom{2}{2}} \frac{p^{b_k+2(k-1)}}{t^{\binom{2}{2}-k}} = p^2$.

Let $h = 3$ so that K_3 is a triangle. Then by Lemma 5.1.3 we know that $g_{K_3}(p, 1/t) =$

$\frac{p^3}{t^2} + \frac{p^5}{t} + p^6$. Now, $\binom{h}{2} = \binom{3}{2} = 3$ and so

$$\sum_{k=1}^{\binom{h}{2}} \frac{p^{b_k+2(k-1)}}{t^{\binom{h}{2}-k}} = \frac{p^{b_1}}{t^2} + \frac{p^{b_2+2}}{t^1} + p^{b_3+4}.$$

Now, b_1 must satisfy $\binom{b_1-1}{2} < \binom{3}{2} - (1-1) = 3 \leq \binom{b_1}{2}$ and so $b_1 = 3$. Similarly, b_2 must satisfy $\binom{b_2-1}{2} < \binom{3}{2} - (2-1) = 2 \leq \binom{b_2}{2}$, and so $b_2 = 3$. Finally, b_3 must satisfy $\binom{b_3-1}{2} < \binom{3}{2} - (3-1) = 1 \leq \binom{b_3}{2}$, and so $b_3 = 2$. So,

$$\sum_{k=1}^{\binom{h}{2}} \frac{p^{b_k+2(k-1)}}{t^{\binom{h}{2}-k}} = \frac{p^{b_1}}{t^2} + \frac{p^{b_2+2}}{t^1} + p^{b_3+4} = \frac{p^3}{t^2} + \frac{p^5}{t^1} + p^6$$

and is the POP of K_3.

Also, for all $h \geq 4$, we have by Theorem 5.2.11 that $g_{K_h}(p, 1/t) = \sum_{k=1}^{\binom{h}{2}} \frac{p^{b_k+2(k-1)}}{t^{\binom{h}{2}-k}}$ is the POP of K_h. Therefore the corollary holds for all $h \geq 2$.

QED.

Theorems 5.2.8 and 5.2.9 and Corollary 5.2.12 state the POP of trees, cycles, and complete graphs. We would like to know the POP of a any graph H. Currently, if given a specific graph H we can calculate its POP by calculating the minimum number of vertices used in graphs induced by a partitions of the edge set (Lemma 5.2.5), however we do not yet have a general form for the POP of H.

5.3 Towards a Threshold Result

In this section we present first moment results for trees, cycles, and complete graphs and a threshold result for K_3. We are able to prove threshold results for small specific graphs, but not in general. However, we are able to describe a general method for proving threshold results if all required POPs are known. The proof of this result, and others in this section, rely heavily on the framework built in Chapter 4.

5.3.1 The First Moment

In this section, we assume that H is a graph on $h > 0$ vertices and $m > 0$ edges. Also, G is a Discrete Random Dot Product Graph drawn from $\mathcal{D}[n, \{0, 1\}^t, p]$. Finally, $\mathbf{Z}_H$ is the number of copies of H in a realization of G.

As first discussed in Section 5.1.2, in any further discussion of the POP, we need to set some conditions on p and t. So, we assume that p and t have the forms $p = \frac{1}{n^\alpha}$ and $t = n^\beta$, where $\alpha, \beta \geq 0$. Then the POP of H will have the form

$$g_H(p, 1/t) = p^{x_0} + \frac{p^{x_1}}{t} + \frac{p^{x_2}}{t^2} + \cdots + \frac{p^{x_{m-2}}}{t^{m-2}} + \frac{p^{x_{m-1}}}{t^{m-1}}$$

$$= (\frac{1}{n^\alpha})^{x_0} + \frac{(\frac{1}{n^\alpha})^{x_1}}{(n^\beta)} + \frac{(\frac{1}{n^\alpha})^{x_2}}{(n^\beta)^2} + \cdots + \frac{(\frac{1}{n^\alpha})^{x_{m-2}}}{(n^\beta)^{m-2}} + \frac{(\frac{1}{n^\alpha})^{x_{m-1}}}{(n^\beta)^{m-1}}$$

$$= \sum_{i=0}^{m-1} \frac{(\frac{1}{n^\alpha})^{x_i}}{(n^\beta)^i}.$$

Furthermore, the expected number of copies of H in a Discrete Random Dot

Product Graph G is

$$E[\mathbf{Z}_H] \asymp \binom{n}{h} g_H(p, 1/t) \asymp n^h \sum_{i=0}^{m-1} \frac{(\frac{1}{n^\alpha})^{x_i}}{(n^\beta)^i} = \sum_{i=0}^{m-1} n^{h-x_i\alpha-\beta i}.$$

Also, by Lemma 5.2.5 and Proposition 5.2.6 we have that $0 < h \leq x_i \leq 2m$ for all $0 \leq i \leq m-1$. So, $E[\mathbf{Z}_H]$ is a posynomial of the form discussed in Chapter 4, with all of the $a_i = 1$ and all of the ordered pairs, (β, α), points in the β, α-parameter space. Therefore, if we let $f^H_{\alpha,\beta,h}(n) = \sum_{i=0}^{m-1} n^{h-x_i\alpha-\beta i}$, then $f^H_{\alpha,\beta,h}(n) \asymp E[\mathbf{Z}_H]$. Proposition 4.1.1 gives that $E[\mathbf{Z}_H] \to 0$ whenever $(\beta, \alpha) \in \mathcal{C}(f^H)$ and $E[\mathbf{Z}_H] \to \infty$ whenever $(\beta, \alpha) \in \zeta(f^H)$. So, we see that $p = \frac{1}{n^\alpha}$, where α is the piecewise linear boundary function $\alpha = F^H(\beta)$, is a candidate for the threshold of the appearance of H in G.

Let us examine this possible threshold function for the classes of graphs for which we can explicitly calculate the POPs. First, we consider the class of trees on $h \geq 2$ vertices.

Proposition 5.3.1. *Let G be drawn from $\mathcal{D}[n, \{0,1\}^t, p]$ with $p = \frac{1}{n^\alpha}$ and $t = n^\beta$ for $\alpha, \beta \geq 0$. Let T be a tree on $h \geq 2$ vertices. Let*

$$f'^T_{\alpha,\beta,h}(n) = n^{h-2(h-1)\alpha} + n^{h-h\,\alpha-\beta\,(h-2)}.$$

Then then the expected number of copies of T in G is

$$E[\mathbf{Z}_T] \equiv f'^T_{\alpha,\beta,h}(n)$$

with $E[\mathbf{Z}_T] \to 0$ *whenever* $(\beta, \alpha) \in \mathcal{C}(f'^T)$ *and* $E[\mathbf{Z}_T] \to \infty$ *whenever* $(\beta, \alpha) \in \zeta(f'^T)$. *Additionally, the boundary function for* $\mathcal{C}(f'^T)$ *is*

$$\alpha = \begin{cases} \frac{h-(h-2)\beta}{h} & \beta \leq \frac{h}{2h-2} \\ \\ \frac{h}{2h-2} & \beta \geq \frac{h}{2h-2} \end{cases}.$$

Proof:

First recall from Theorem 5.2.8 that the POP of the tree T is

$$g_T(p, \frac{1}{t}) = p^{2(h-1)} + \frac{p^{2(h-1)-1}}{t} + \cdots + \frac{p^{h+1}}{t^{h-3}} + \frac{p^h}{t^{h-2}} = \sum_{i=0}^{h-2} \frac{p^{2(h-1)-i}}{t^i}.$$

Therefore the expected number of copies of T in G is

$$E[\mathbf{Z}_T] \asymp \binom{n}{h} g_T(p, 1/t) \asymp f^T_{\alpha,\beta,h}(n) = \sum_{i=0}^{h-2} n^{h-(2(h-1)-i)\alpha-\beta i}$$

$$= n^{h-2(h-1)\alpha} + \left(\sum_{i=1}^{h-3} n^{h-(2(h-1)-i)\alpha-\beta i}\right) + n^{h-h\alpha-\beta(h-2)}$$

$$= f'^T_{\alpha,\beta,h}(n) + \sum_{i=1}^{h-3} n^{h-(2(h-1)-i)\alpha-\beta i}.$$

If we show that $f^T_{\alpha,\beta,h}(n) \equiv f'^T_{\alpha,\beta,h}(n)$ then we will be done.

In the context of Chapter 4, for each $0 \leq i \leq h-2$ let l_i be the line

$$h - (2(h-1) - i)\alpha - \beta i = 0.$$

Then by Proposition 4.1.1, $\mathcal{C}(f^T)$ is determined by $l_0, \ldots, l_{h-2}$ and $\mathcal{C}(f^{'T})$ is determined by l_0 and l_{h-2}. If we can show that $\mathcal{C}(f^T)$ is determined by just l_0 and l_{h-2}, then clearly $\mathcal{C}(f^T) = \mathcal{C}(f^{'T})$ and hence by Proposition 4.2.2 we will have that $f^T_{\alpha,\beta,h}(n) \equiv f^{'T}_{\alpha,\beta,h}(n)$. Recall that by Proposition 4.2.3, to show that $\mathcal{C}(f^T)$ is determined by just l_0 and l_{h-2}, it is sufficient to show that for each $i \in \{1, \ldots, h-3\}$, l_i falls 'below' $\mathcal{C}(f^{'T})$, the convex set determined by l_0 and l_{h-2}.

Now, l_0 is the horizontal line $h - 2(h-1)\alpha = 0$ and has α-intercept $\alpha_0 = \frac{h}{h-2}$. Also, l_{h-2} is the line $h - h\alpha - \beta\,(h-2) = 0$ and has α-intercept $\alpha_{h-2} = 1$. Finally, l_0 and l_{h-2} intersect within the β,α-parameter space at the point $(\frac{h}{2h-2}, \frac{h}{2h-2})$.

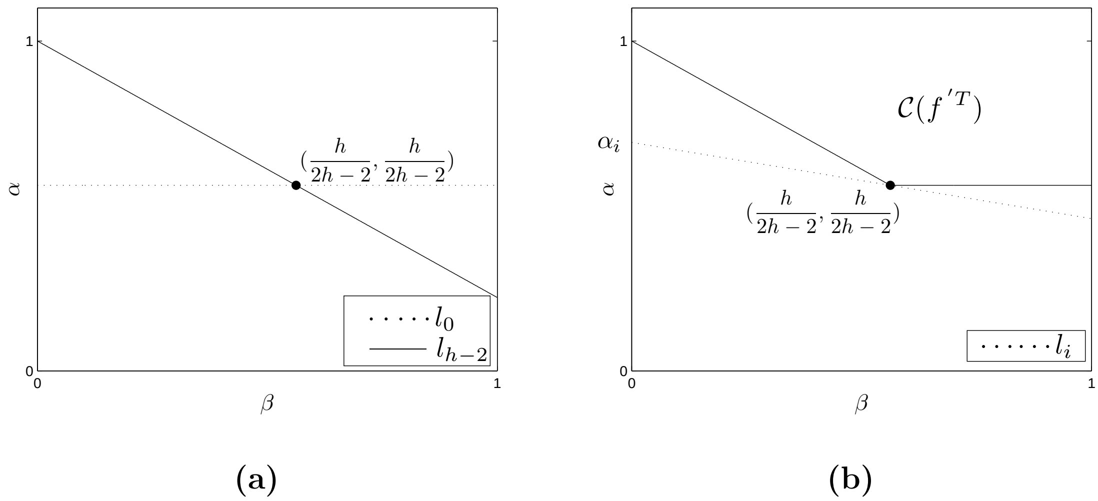


Figure 5.1: **(a)** The lines that define $\mathcal{C}(f^{'T})$. **(b)** l_i intersects $\mathcal{C}(f^T)$ at exactly one point.

Now, let $i \in \{1, \ldots, h-3\}$ we wish to show that $l_i : h - (2(h-1) - i)\alpha - \beta i = 0$ falls 'below' $\mathcal{C}(f'^T)$. The line l_i has α-intercept $\alpha_i = \frac{h}{2h-2-i}$. Also, $i \leq h-3$, implying that $2h - 2 - i \geq 2h - 2 - h + 3 = h + 1 > h$. Therefore, $\alpha_i = \frac{h}{2h-2-i} < 1 = \alpha_{h-2}$ and the α-intercept of l_i is below that of l_{h-2}. Next, it is easy to see that l_i also goes through point $(\frac{h}{2h-2}, \frac{h}{2h-2})$, the intersection point of l_0 and l_{h-2}. Finally, l_i has negative slope, $-\frac{i}{2(h-1)-i}$. Hence, l_i intersects the boundary of $\mathcal{C}(f^T)$ at exactly one point.

So, by Propositions 4.2.3 and 4.2.5 we see that $f'^T_{\alpha,\beta,h}(n)$ can be obtained from $f^T_{\alpha,\beta,h}(n)$ by removing each of the terms associated with the lines l_i, $1 \leq i \leq h-3$, and furthermore $f'^T_{\alpha,\beta,h}(n) = f^{red\,T}_{\alpha,\beta,h}(n)$. Hence $f'^T_{\alpha,\beta,h}(n) \equiv f^T_{\alpha,\beta,h}(n)$ and so $E[\mathbf{Z}_T] \equiv f'^T_{\alpha,\beta,h}(n)$ with $E[\mathbf{Z}_T] \to 0$ whenever $(\beta, \alpha) \in \mathcal{C}(f'^T)$ and $E[\mathbf{Z}_T] \to \infty$ whenever $(\beta, \alpha) \in \zeta(f'^T)$.

QED.

Therefore, by Markov's inequality $P[\mathbf{Z}_T \geq 1] \leq E[\mathbf{Z}_T] \to 0$ as $n \to \infty$, whenever $(\beta, \alpha) \in \mathcal{C}(f'^T)$ and almost no graph contains T. Additionally, $p = \frac{1}{n^\alpha}$, where α is the piecewise linear boundary of $\mathcal{C}(f'^T)$,

$$\alpha = \begin{cases} \frac{h-(h-2)\beta}{h} & \beta \leq \frac{h}{2h-2} \\ \\ \frac{h}{2h-2} & \beta \geq \frac{h}{2h-2} \end{cases},$$

is a candidate for the threshold for the appearance of T.

Next, we consider the cycles on $h \geq 3$ vertices.

Proposition 5.3.2. *Let G be drawn from $\mathcal{D}[n, \{0,1\}^t, p]$ with $p = \frac{1}{n^\alpha}$ and $t = n^\beta$ for $\alpha, \beta \geq 0$. Let C_h be a cycle on $h \geq 3$ vertices. Let*

$$f'^{C_h}_{\alpha,\beta,h}(n) = n^{h-2h\alpha} + n^{h-h\,\alpha-\beta\,(h-1)}.$$

Then then the expected number of copies of C_h in G is

$$E[\mathbf{Z}_{C_h}] \equiv f'^{C_h}_{\alpha,\beta,h}(n)$$

with $E[\mathbf{Z}_{C_h}] \to 0$ whenever $(\beta, \alpha) \in \mathcal{C}(f'^{C_h})$ and $E[\mathbf{Z}_{C_h}] \to \infty$ whenever $(\beta, \alpha) \in \zeta(f'^{C_h})$. Additionally, the boundary function for $\mathcal{C}(f'^{C_h})$ is

$$\alpha = \begin{cases} \frac{h-(h-1)\beta}{h} & \beta \leq \frac{h}{2h-2} \\ \\ \frac{1}{2} & \beta \geq \frac{h}{2h-2} \end{cases}.$$

Proof:

First recall from Theorem 5.2.9 that the POP of C_h is

$$g_{C_h}(p, \frac{1}{t}) = p^{2h} + \frac{p^{2h-1}}{t} + \cdots + \frac{p^{2h+2}}{t^{h-2}} + \frac{p^h}{t^{h-1}} = \sum_{i=2}^{h} \frac{p^{h+i}}{t^{h-i}} + \frac{p^h}{t^{h-1}}.$$

Therefore the expected number of copies of C_h in G is

$$E[\mathbf{Z}_{C_h}] \asymp \binom{n}{h} g_{C_h}(p, 1/t) \asymp f^{C_h}_{\alpha,\beta,h}(n) = \sum_{i=2}^{h} n^{h-(h+i)\alpha-\beta(h-i)} + n^{h-h\alpha-\beta(h-1)}$$

$$= n^{h-2h\alpha} + \left(\sum_{i=2}^{h-1} n^{h-(h+i)\alpha-\beta(h-i)} \right) + n^{h-h\alpha-\beta(h-1)}$$

$$= f'^{C_h}_{\alpha,\beta,h}(n) + \sum_{i=2}^{h-1} n^{h-(h+i)\alpha-\beta(h-i)}.$$

If we show that $f^{C_h}_{\alpha,\beta,h}(n) \equiv f'^{C_h}_{\alpha,\beta,h}(n)$ then we are done.

In the context of Chapter 4, let l_1 be the line $h - h\alpha - \beta(h-1) = 0$. For each $2 \le i \le h$, let l_i be the line

$$h - (h+i)\alpha - \beta(h-i) = 0.$$

Then by Proposition 4.1.1, $\mathcal{C}(f^{C_h})$ is determined by $l_1, \ldots, l_h$ and $\mathcal{C}(f'^{C_h})$ is determined by l_1 and l_h. If we can show that $\mathcal{C}(f^{C_h})$ is determined by just l_1 and l_h, then clearly $\mathcal{C}(f^{C_h}) = \mathcal{C}(f'^{C_h})$ and hence by Proposition 4.2.2 we will have that $f^{C_h}_{\alpha,\beta,h}(n) \equiv f'^{C_h}_{\alpha,\beta,h}(n)$. Recall that by Proposition 4.2.3, to show that $\mathcal{C}(f'^{C_h})$ is determined by l_1 and l_h, it is sufficient to show that for each $i \in \{2, \ldots, h-1\}$, l_i falls 'below' $\mathcal{C}(f'^{C_h})$, the convex set determined by l_1 and l_h.

Now, l_h is the horizontal line $h - 2h\alpha = 0$ and has α-intercept $\alpha_h = \frac{1}{2}$. Also, l_1 is the line $h - h\alpha - \beta(h-1) = 0$ and has α-intercept $\alpha_1 = 1$. Finally, l_1 and l_h intersect

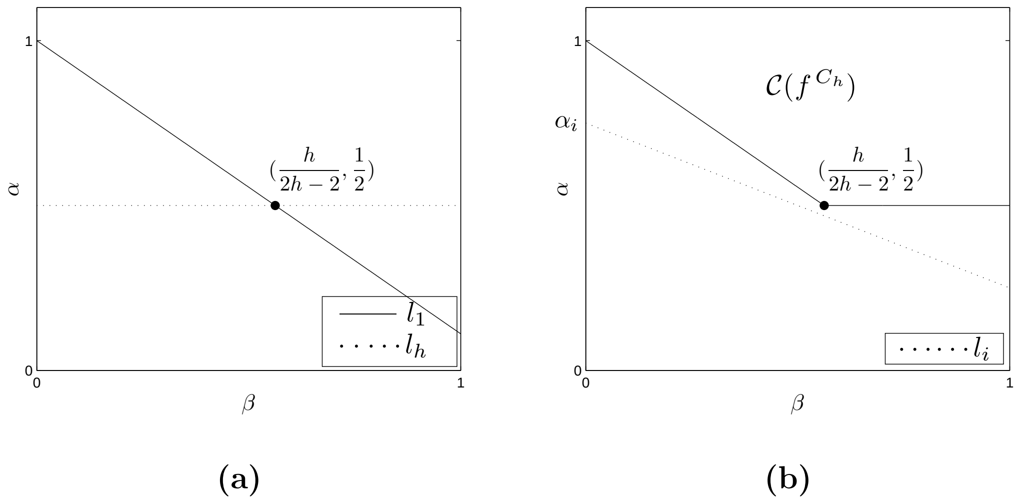


Figure 5.2: **(a)** The lines that define $\mathcal{C}(f'^{C_h})$. **(b)** l_i is 'below' $\mathcal{C}(f^{C_h})$.

within the β, α-parameter space at the point $(\frac{h}{2h-2}, \frac{1}{2})$.

Now, let $i \in \{2, \ldots, h-1\}$ we wish to show that $l_i : h - (h+i)\alpha - \beta(h-i) = 0$ falls 'below' $\mathcal{C}(f'^{C_h})$. The line l_i has α-intercept $\alpha_i = \frac{h}{h+i}$. Also, $i \geq 2$, implying that $h + i \geq h + 2 > h$. Therefore, $\alpha_i = \frac{h}{h+i} < 1 = \alpha_1$ and the α-intercept of l_i is below that of l_1. Next, it is easy to see that l_i goes through point $(\frac{h}{2h-2}, \frac{h^2-2h+hi}{2(h^2-h-k+hi)})$. Now, we know that $i < h$ therefore

$$-2h < -h - i$$

and

$$h^2 - 2h + hi < h^2 - h - i + hi$$

and so $\frac{h^2-2h+hi}{2(h^2-h-i+hi)} < \frac{1}{2}$. So, at $\beta = \frac{h}{2h-2}$, l_i is under the intersection point of l_1 and l_h and hence below $\mathcal{C}(f'^{C_h})$. Additionally, the slope of l_i is $\frac{-(h-i)}{h+i}$ and is negative since $h - i \geq h - (h-1) \geq 1$. These two facts, along with the facts that $\alpha_i < \alpha_1$ and l_h is a horizontal line, guarantee that l_i is 'below' $\mathcal{C}(f'^{C_h})$ for all $2 \leq i \leq h-1$.

So, by Propositions 4.2.3 and 4.2.5 we see that $f'^{C_h}_{\alpha,\beta,h}(n)$ can be obtained from $f^{C_h}_{\alpha,\beta,h}(n)$ by removing each of the terms associated with the lines l_i, $2 \leq i \leq h-1$, and furthermore $f'^{C_h}_{\alpha,\beta,h}(n) = f^{red\,C_h}_{\alpha,\beta,h}(n)$. Hence $f'^{C_h}_{\alpha,\beta,h}(n) \equiv f^{C_h}_{\alpha,\beta,h}(n)$ and so $E[\mathbf{Z}_{C_h}] \equiv f'^{C_h}_{\alpha,\beta,h}(n)$ with $E[\mathbf{Z}_{C_h}] \to 0$ whenever $(\beta,\alpha) \in \mathcal{C}(f'^{C_h})$ and $E[\mathbf{Z}_{C_h}] \to \infty$ whenever $(\beta,\alpha) \in \zeta(f'^{C_h})$.

QED.

Therefore, by Markov's inequality $P[\mathbf{Z}_{C_h} \geq 1] \leq E[\mathbf{Z}_{C_h}] \to 0$, whenever $(\beta,\alpha) \in \mathcal{C}(f'^{C_h})$ as $n \to \infty$ and almost no graph contains C_h. Additionally, $p = \frac{1}{n^\alpha}$, where α is the piecewise linear boundary of $\mathcal{C}(f'^{C_h})$,

$$\alpha = \begin{cases} \frac{h-(h-1)\beta}{h} & \beta \leq \frac{h}{2h-2} \\ & \\ \frac{1}{2} & \beta \geq \frac{h}{2h-2} \end{cases},$$

is a candidate for the threshold for the appearance of C_h.

Now, let us consider the complete graph K_h on $h \geq 3$ vertices.

Proposition 5.3.3. *Let G be drawn from $\mathcal{D}[n,\{0,1\}^t,p]$ with $p = \frac{1}{n^\alpha}$ and $t = n^\beta$ for $\alpha,\beta \geq 0$. Let K_h be the complete graph on $h \geq 3$ vertices. Let*

$$f'^{K_h}_{\alpha,\beta,h}(n) = n^{h-h\alpha-\beta(\binom{h}{2}-1)} + n^{h-2\binom{h}{2}\alpha}.$$

Then then the expected number of copies of K_h in G is

$$E[\mathbf{Z}_{K_h}] \equiv f'^{K_h}_{\alpha,\beta,h}(n)$$

with $E[\mathbf{Z}_{K_h}] \to 0$ whenever $(\beta,\alpha) \in \mathcal{C}(f'^{K_h})$ and $E[\mathbf{Z}_{K_h}] \to \infty$ whenever $(\beta,\alpha) \in \zeta(f'^{K_h})$. Additionally, the boundary function for $\mathcal{C}(f'^{K_h})$ is

$$\alpha = \begin{cases} \frac{h-(\binom{h}{2}-1)\beta}{h} & \beta \leq \frac{2h}{h^2-1} \\ \\ \frac{1}{h-1} & \beta \geq \frac{2h}{h^2-1} \end{cases}.$$

Proof:

First recall from Corollary 5.2.12 that the POP of K_h is

$$g_{K_h}(p, 1/t) = \sum_{k=1}^{\binom{h}{2}} \frac{p^{b_k+2(k-1)}}{t^{\binom{h}{2}-k}},$$

where for any $1 \leq k \leq \binom{h}{2}$, $b_k \in \mathbb{Z}_+$ with $\binom{b_k-1}{2} < \binom{h}{2} - (k-1) \leq \binom{b_k}{2}$. Therefore the expected number of copies of K_h in G is

$$E[\mathbf{Z}_{K_h}] \asymp \binom{n}{h} g_{K_h}(p, 1/t) \asymp f^{K_h}_{\alpha,\beta,h}(n) = \sum_{k=1}^{\binom{h}{2}} n^{h-(b_k+2(k-1))\alpha-\beta(\binom{h}{2}-k)}.$$

Now, when $k = 1$, b_1 must satisfy $\binom{b_1-1}{2} < \binom{h}{2} \leq \binom{b_1}{2}$. Hence, $b_1 = h$. Also, if $k = \binom{h}{2}$, then $b_{\binom{h}{2}}$ must satisfy $\binom{b_k-1}{2} < 1 \leq \binom{b_k}{2}$ and so $b_{\binom{h}{2}} = 2$. So, we can rewrite

$f^{K_h}_{\alpha,\beta,h}(n)$ as

$$f^{K_h}_{\alpha,\beta,h}(n) = n^{h-h\alpha-\beta(\binom{h}{2}-1)} + \left(\sum_{k=2}^{\binom{h}{2}-1} n^{h-(b_k+2(k-1))\alpha-\beta(\binom{h}{2}-k)}\right) + n^{h-2\binom{h}{2}\alpha}$$

$$= f'^{K_h}_{\alpha,\beta,h}(n) + \sum_{k=2}^{\binom{h}{2}-1} n^{h-(b_k+2(k-1))\alpha-\beta(\binom{h}{2}-k)}.$$

If we show that $f^{K_h}_{\alpha,\beta,h}(n) \equiv f'^{K_h}_{\alpha,\beta,h}(n)$ then we will be done.

In the context of Chapter 4, for each $1 \leq k \leq \binom{h}{2}$ let l_k be the line

$$h - (b_k + 2(k-1))\alpha - \beta(\binom{h}{2} - k) = 0.$$

Then by Proposition 4.1.1, $\mathcal{C}(f^{K_h})$ is determined by $l_1, \ldots, l_{\binom{h}{2}}$ and $\mathcal{C}(f'^{K_h})$ is determined by l_1 and $l_{\binom{h}{2}}$. If we can show that $\mathcal{C}(f^{K_h})$ is determined by just l_1 and $l_{\binom{h}{2}}$, then clearly $\mathcal{C}(f^{K_h}) = \mathcal{C}(f'^{K_h})$ and hence by Proposition 4.2.2 we will have that $f^{K_h}_{\alpha,\beta,h}(n) \equiv f'^{K_h}_{\alpha,\beta,h}(n)$. Recall that by Proposition 4.2.3, to show that $\mathcal{C}(f^{K_h})$ is determined by just l_1 and $l_{\binom{h}{2}}$, it is sufficient to show that for each $k \in \{2, \ldots, \binom{h}{2} - 1\}$, l_k falls 'below' $\mathcal{C}(f'^{K_h})$, the convex set determined by l_1 and $l_{\binom{h}{2}}$.

Now, l_1 is the line $h - h\alpha - \beta(\binom{h}{2} - 1) = 0$ and has α-intercept $\alpha_1 = 1$ and β-intercept $\beta_1 = \frac{h}{\binom{h}{2}-1}$. Also, line $l_{\binom{h}{2}}$ is the horizontal line $h - 2\binom{h}{2}\alpha = 0$ and has α-intercept $\alpha_{\binom{h}{2}} = \frac{h}{2\binom{h}{2}} = \frac{1}{h-1}$. Finally, l_1 and $l_{\binom{h}{2}}$ intersect within the β, α-parameter space at the point $(\frac{2h}{h^2-1}, \frac{1}{h-1})$.

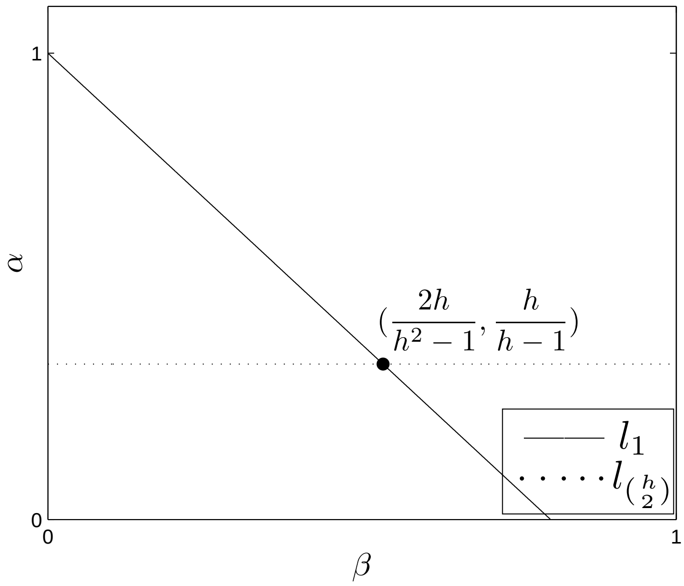


Figure 5.3: The lines that define $\mathcal{C}(f'^{K_h})$.

Now, let $k \in \{2, \ldots, \binom{h}{2} - 1\}$ we wish to show that $l_k : h - (b_k + 2(k-1))\alpha - \beta(\binom{h}{2} - k) = 0$ falls 'below' $\mathcal{C}(f'^{K_h})$. First note that $b_k + 2(k-1) > 0$ and $\binom{h}{2} - k > 0$, therefore l_k has the negative slope $-\frac{b_k+2(k-1)}{\binom{h}{2}-k}$. Also, l_k has α-intercept $\alpha_k = \frac{h}{b_k+2(k-1)}$.

Claim 1: $\alpha_k = \frac{h}{b_k+2(k-1)} < 1 = \alpha_1$

Proof of Claim 1: By assumption, we have that

$$\binom{h}{2} - (k-1) \leq \binom{b_k}{2} = \frac{b_k(b_k-1)}{2}.$$

Therefore,

$$\frac{h(h-1)}{2} \leq \frac{b_k(b_k-1)}{2} + (k-1)$$

and

$$h(h-1) \leq b_k(b_k-1) + 2(k-1).$$

Now, $h \geq b_k$ since $\binom{b_k-1}{2} < \binom{h}{2} - (k-1)$, so

$$h(h-1) \leq b_k(h-1) + 2(k-1).$$

Also, $h-1 > 1$, hence

$$h(h-1) < b_k(h-1) + 2(k-1)(h-1)$$

and so we see that

$$\frac{h}{b_k + 2(k-1)} < 1.$$

Hence, $\alpha_k < \alpha_1$ and Claim 1 holds.

So, we have that l_k has negative slope and α-intercept below that of l_1. Since l_k is not a vertical line, we know that l_k goes thought the point $(\frac{2h}{h^2-1}, \alpha_0)$ for some α_0. If we can show that $\alpha_0 < \frac{1}{h-1}$, then the point $(\frac{2h}{h^2-1}, \alpha_0)$ is directly under the intersection point of l_1 and $l_{\binom{h}{2}}$, $(\frac{2h}{h^2-1}, \frac{1}{h-1})$.

Claim 2: l_k a goes though the point $(\frac{2h}{h^2-1}, \alpha_0)$ for some $\alpha_0 < \frac{1}{h-1}$

Proof of Claim 2: The line l_k is the line $h - (b_k + 2(k-1))\alpha - \beta(\binom{h}{2} - k) = 0$, therefore when $\beta = \frac{2h}{h^2-1}$ we have that

$$h - (b_k + 2(k-1))\,\alpha_0 - \frac{2h}{h^2-1}\left(\binom{h}{2} - k\right) = 0$$

and it is easy to see that

$$\alpha_0 = \frac{2kh}{(b_k + 2(k-1))(h+1)(h-1)}.$$

Now, $\frac{h}{h+1} < 1$ and so

$$\alpha_0 < \frac{2k}{(b_k + 2(k-1))(h-1)}.$$

Now, by assumption, $\binom{b_k}{2} \geq \binom{h}{2} - (k-1) \geq 1$, therefore $b_k \geq 2$ and $b_k + 2(k-1) \geq 2k$. Hence $\frac{2k}{b_k+2(k-1)} \leq 1$ and so we see that

$$\alpha_0 < \frac{1}{h-1}$$

and Claim 2 holds.

Claims 1 and 2 together with the facts that l_k has negative slope and $l_{\binom{h}{2}}$ is a horizontal line, ensures that l_k is 'below' $\mathcal{C}(f'^{K_h})$ for all $k \in \{2, \ldots, \binom{h}{2} - 1\}$.

Hence by Propositions 4.2.3 and 4.2.5 we see that $f'^{K_h}_{\alpha,\beta,h}(n)$ can be obtained from $f^{K_h}_{\alpha,\beta,h}(n)$ by removing each of the terms associated with the lines l_k, $2 \leq k \leq \binom{h}{2} - 1$, and furthermore $f'^{K_h}_{\alpha,\beta,h}(n) = f^{red\,K_h}_{\alpha,\beta,h}(n)$. Hence $f'^{K_h}_{\alpha,\beta,h}(n) \equiv f^{K_h}_{\alpha,\beta,h}(n)$ and so $E[\mathbf{Z}_{K_h}] \equiv f'^{K_h}_{\alpha,\beta,h}(n)$ with $E[\mathbf{Z}_{K_h}] \to 0$ whenever $(\beta, \alpha) \in \mathcal{C}(f'^{K_h})$ and $E[\mathbf{Z}_{K_h}] \to \infty$ whenever

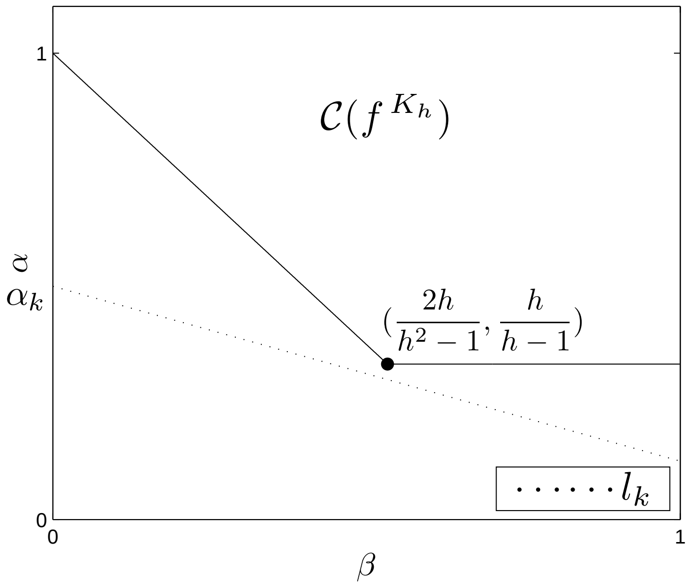


Figure 5.4: The line l_k is 'below' $\mathcal{C}(f^{C_h})$.

$(\beta, \alpha) \in \zeta(f'^{K_h})$.

QED.

Therefore $\frac{1}{n^\alpha}$, where α is the piecewise linear boundary of $\mathcal{C}(f'^{K_h})$,

$$\alpha = \begin{cases} \frac{h-(\binom{h}{2}-1)\beta}{h} & \beta \leq \frac{2h}{h^2-1} \\ \\ \frac{1}{h-1} & \beta \geq \frac{2h}{h^2-1} \end{cases},$$

is a candidate for the threshold for the appearance of K_h.

In the three classes of graphs that we have investigated so far, the expected number of graphs H has always been associated with convex sets (in which $E[Z_H] \to 0$) that are determined by only the first and last terms in the POP of H. This might lead one to conjecture that this is always the case. However, we show that this conjecture is false.

Let H_{kite} be the graph with vertex set $V(H_{kite}) = \{v_1, v_2, v_3, v_4, v_5\}$ and edge set

$$E(H_{kite}) = \{\{v_1, v_2\}, \{v_1, v_3\}, \{v_1, v_4\}, \{v_1, v_5\}, \{v_2, v_3\}, \{v_2, v_4\}, \{v_3, v_4\}\}.$$

Since H_{kite} is a K_4 with a single pendant edge we call H_{kite} the 'kite' graph. See Figure 5.5.

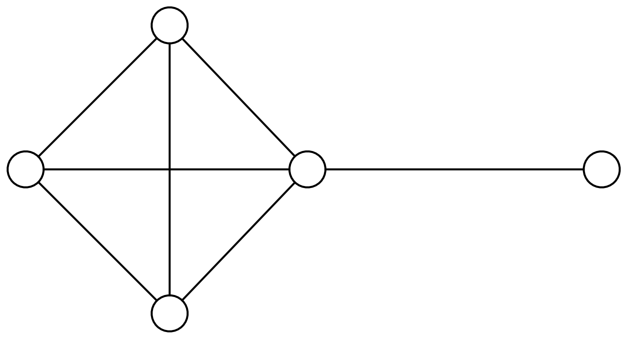


Figure 5.5: The graph H_{kite}.

Proposition 5.3.4. *Let G be drawn from $\mathcal{D}[n, \{0,1\}^t, p]$ with $p = \frac{1}{n^\alpha}$ and $t = n^\beta$ for $\alpha, \beta \geq 0$. Let*

$$f'^{H_{kite}}_{\alpha,\beta,5}(n) = n^{5-14\alpha} + n^{5-6\alpha-5\beta} + n^{5-5\alpha-6\beta}.$$

Then

- *the POP of H_{kite} is $g_{H_{kite}}(p, 1/t) = p^{14} + \frac{p^{13}}{t} + \frac{p^{11}}{t^2} + \frac{p^{10}}{t^3} + \frac{p^8}{t^4} + \frac{p^6}{t^5} + \frac{p^5}{t^6}$,*
- *the expected number of copies of H_{kite} in G is $E[\mathbf{Z}_{H_{kite}}] \equiv f'^{H_{kite}}_{\alpha,\beta,5}(n)$ with*

$E[\mathbf{Z}_{H_{kite}}] \to 0$ whenever $(\beta,\alpha) \in \mathcal{C}({f'}^{H_{kite}})$ and $E[\mathbf{Z}_{H_{kite}}] \to \infty$ whenever $(\beta,\alpha) \in \zeta({f'}^{H_{kite}})$, and

- $E[\mathbf{Z}_{H_{kite}}] \not\asymp n^{5-14\alpha} + n^{5-5\alpha-6\beta}$.

Additionally, the boundary function for $\mathcal{C}({f'}^{H_{kite}})$ is

$$\alpha = \begin{cases} \frac{5-6\beta}{5} & \beta \leq \frac{5}{11} \\ \frac{5-5\beta}{6} & \frac{5}{11} \leq \beta \leq \frac{4}{7} \\ \frac{5}{14} & \beta \geq \frac{4}{7} \end{cases}.$$

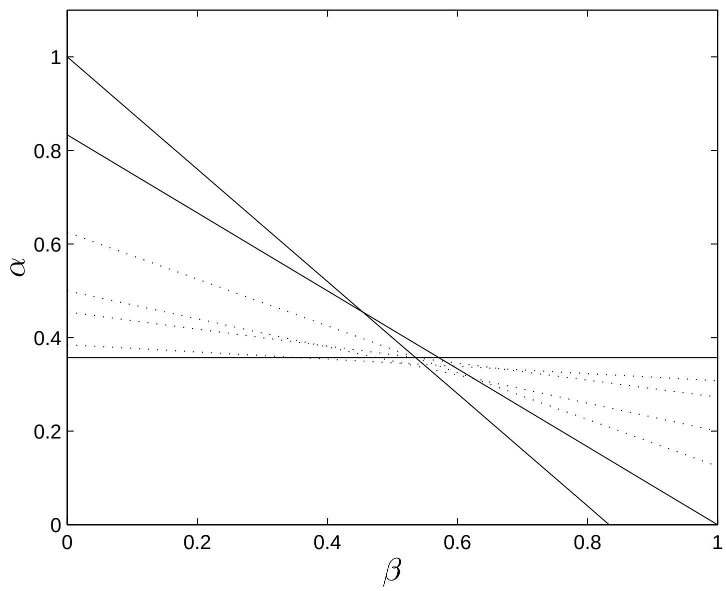
Proof:

First, it is easy to show, either directly using Lemma 5.2.5 or by calculating the POP of K_4 by Corollary 5.2.12 and then applying the pendant edge Lemma 5.2.7, that the POP of H_{kite} is $g_{H_{kite}}(p,1/t) = p^{14} + \frac{p^{13}}{t} + \frac{p^{11}}{t^2} + \frac{p^{10}}{t^3} + \frac{p^8}{t^4} + \frac{p^6}{t^5} + \frac{p^5}{t^6}$. Therefore, the expected number of copies of H_{kite} is G is

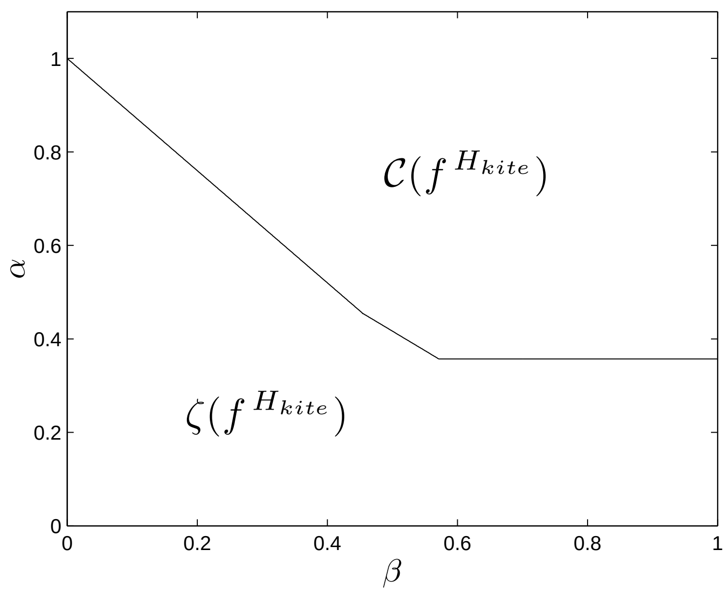
$$E[\mathbf{Z}_{H_{kite}}] \asymp \binom{n}{5} g_{H_{kite}}(p,1/t) \asymp f^{H_{kite}}_{\alpha,\beta,5}(n)$$

$$= n^{5-14\alpha} + n^{5-13\alpha-\beta} + n^{5-11\alpha-2\beta} + n^{5-10\alpha-3\beta} + n^{5-8\alpha-4\beta} + n^{5-6\alpha-5\beta} + n^{5-5\alpha-6\beta}.$$

Now, in Figure 5.6 (a) we plot each of the lines associated with the terms in



(a)



(b)

Figure 5.6: **(a)** All of the lines associated with terms of $f^{H_{kite}}$. **(b)** $\mathcal{C}(f^{H_{kite}})$ and $\zeta(f^{H_{kite}})$.

$f^{H_{kite}}_{\alpha,\beta,5}(n)$ in the β,α-parameter space. Clearly, the convex region $\mathcal{C}(f^{H_{kite}})$ is determined by the solid lines associated with the first, last and penultimate terms and that the dotted lines associated with all of the other terms fall 'below' $\mathcal{C}(f^{H_{kite}})$.

Hence by Propositions 4.2.3 and 4.2.5 we see that $f'^{H_{kite}}_{\alpha,\beta,5}(n)$ can be obtained from $f^{H_{kite}}_{\alpha,\beta,5}(n)$ by removing each of the terms associated with the dashed lines and furthermore $f'^{K_h}_{\alpha,\beta,h}(n) = f^{red\,K_h}_{\alpha,\beta,h}(n)$. Therefore

$$E[\mathbf{Z}_{H_{kite}}] \asymp f^{H_{kite}}_{\alpha,\beta,5}(n) \equiv f'^{H_{kite}}_{\alpha,\beta,5}(n) = n^{5-14\alpha} + n^{5-6\alpha-5\beta} + n^{5-5\alpha-6\beta}$$

and hence $\mathcal{C}(f^{H_{kite}}) = \mathcal{C}(f'^{H_{kite}})$. Additionally, since $f'^{K_h}_{\alpha,\beta,h}(n) = f^{red\,K_h}_{\alpha,\beta,h}(n)$, $E[\mathbf{Z}_{H_{kite}}] \not\equiv n^{5-14\alpha} + n^{5-5\alpha-6\beta}$.

QED.

So the convex set associated with the expected number of copies of H_{kite} has a piecewise linear boundary composed of three pieces and not just the two conjectured.

5.3.2 Threshold Theory

We begin this section by presenting a threshold result for K_3.

First, recall that by Corollary 5.2.12 that the POP of K_3 is $g_{K_3}(p, 1/t) = p^6 + \frac{p^5}{t} + \frac{p^3}{t^2}$. Also, recall from Proposition 5.3.3 that the expected number of copies of K_3 goes to zero for all $(\beta,\alpha) \in \mathcal{C}(f'^{K_3})$ and goes to infinity for all $(\beta,\alpha) \in \zeta(f'^{K_3})$, where $f'^{K_3}_{\alpha,\beta,3}(n) = n^{3-3\alpha-2\beta} + n^{3-6\alpha}$.

Theorem 5.3.5. *Let G be drawn from $\mathcal{D}[n, \{0,1\}^t, p]$ with $p = \frac{1}{n^\alpha}$ and $t = n^\beta$ for $\alpha, \beta \geq 0$. Let*

$$f'^{K_3}_{\alpha,\beta,3}(n) = n^{3-3\alpha-2\beta} + n^{3-6\alpha}.$$

Then a threshold function for the appearance of K_3 in G is $p = \frac{1}{n^\alpha}$, where α is the boundary of the convex region $\mathcal{C}(f'^{K_3})$,

$$\alpha = F'^{K_3}(\beta) = \begin{cases} \frac{3-2\beta}{3} & \beta \leq \frac{3}{4} \\ \\ \frac{1}{2} & \beta \geq \frac{3}{4} \end{cases}.$$

That is, with high probability G will be contain no copies of K_3 whenever $(\beta, \alpha) \in \mathcal{C}(f'^{K_3})$ and with high probability G will have at least one K_3 whenever $(\beta, \alpha) \in \zeta(f'^{K_3})$.

Proof:

Let the random variable $\mathbf{Z}_{K_3}$ be the number triangles in G. Then we know from Proposition 5.3.3 that $E[\mathbf{Z}_{K_3}] \equiv f'^{K_3}_{\alpha,\beta,3}(n) = n^{3-3\alpha-2\beta} + n^{3-6\alpha}$. Let $p = \frac{1}{n^\alpha}$, where $\alpha = F'^{K_3}(\beta)$, be our candidate for the threshold for the appearance of triangles. In other words we wish to show that with high probability G will be contain no copies of K_3 whenever $(\beta, \alpha) \in \mathcal{C}(f'^{K_3})$ and with high probability G will have at least one K_3 whenever $(\beta, \alpha) \in \zeta(f'^{K_3})$. Now, by Markov's inequality $P[\mathbf{Z}_{K_3} \geq 1] \leq E[\mathbf{Z}_{K_3}] \to 0$ as $n \to \infty$ whenever $(\beta, \alpha) \in \mathcal{C}(f'^{K_3})$, since $E[\mathbf{Z}_{K_3}] \equiv f'^{K_3}_{\alpha,\beta,3}(n)$. Therefore the

$P[\mathbf{Z}_{K_3} = 0] \to 1$ as $n \to \infty$ and almost every graph has no K_3.

Also, whenever $(\beta, \alpha) \in \zeta(f'^{K_3})$, then $E[\mathbf{Z}_{K_3}] \to \infty$ and we use the second moment method. By Chebychev's inequality we know that $P[\mathbf{X} = 0] \leq \mathrm{Var}(\mathbf{Z}_{K_3})/E[\mathbf{Z}_{K_3}]^2 = \frac{E[\mathbf{Z}_{K_3}^2] - E[\mathbf{Z}_{K_3}]^2}{E[\mathbf{Z}_{K_3}]^2}$.

Now, in Section 5.1.2 we showed if $a, b, c, u, v, w \in V(G)$, $u < v < w$, $a < b < c$, all distinct vertices, that the first term in the variance calculation is

$$E[\mathbf{Z}^2] = \binom{n}{3} P[\text{the triangle } uvw] + \binom{n}{3}(n-3)3P[\{\text{the triangle } uvw\} \cap \{\text{the triangle } avw\}]$$

$$+ \binom{n}{3}\binom{n-3}{2} 3P[\{\text{the triangle } uvw\} \cap \{\text{the triangle } abw\}]$$

$$+ \binom{n}{3}\binom{n-3}{3} P[\{\text{the triangle } uvw\} \cap \{\text{the triangle } abc\}].$$

Now, using Lemma 5.2.5 to calculate the POPs of the required graphs in the equation we see that

$$E[\mathbf{Z}^2] \asymp E[\mathbf{Z}_{k_3}] + n^4(p^{10} + \frac{p^9}{t} + \frac{p^7}{t^2} + \frac{p^6}{t^3} + \frac{p^4}{t^4}) + n^5(p^{12} + \frac{p^{11}}{t} + \frac{p^9}{t^2} + \frac{p^8}{t^3} + \frac{p^6}{t^4} + \frac{p^5}{t^5})$$

$$+ \binom{n}{3}\binom{n-3}{3}(P[\text{the triangle } uvw])^2.$$

Now, let

$$g^*_{\alpha,\beta,4}(n) = n^{4-10\alpha} + n^{4-9\alpha-\beta} + n^{4-7\alpha-2\beta} + n^{4-6\alpha-3\beta} + n^{4-4\alpha-4\beta}$$

and

$$g^{**}_{\alpha,\beta,5}(n) = n^{5-12\alpha} + n^{5-11\alpha-\beta} + n^{5-9\alpha-2\beta} + n^{5-8\alpha-3\beta} + n^{5-6\alpha-4\beta} + n^{5-5\alpha-5\beta}.$$

Then we have

$$E[\mathbf{Z}^2] \asymp E[\mathbf{Z}_{k_3}] + g^{*}_{\alpha,\beta,4}(n) + g^{**}_{\alpha,\beta,5}(n) + \binom{n}{3}\binom{n-3}{3}(P[\text{the triangle } uvw])^2.$$

And by Chebychev's we see that

$$P[\mathbf{Z}_{K_3} = 0] \leq \frac{E[\mathbf{Z}^2_{K_3}] - E[\mathbf{Z}_{K_3}]^2}{E[\mathbf{Z}_{K_3}]^2}$$

$$\asymp \frac{E[\mathbf{Z}_{k_3}] + g^{*}_{\alpha,\beta,4}(n) + g^{**}_{\alpha,\beta,5}(n) + \binom{n}{3}\binom{n-3}{3}(P[\text{the triangle } uvw])^2 - E[\mathbf{Z}_{K_3}]^2}{E[\mathbf{Z}_{K_3}]^2}$$

$$= \frac{E[\mathbf{Z}_{k_3}]}{E[\mathbf{Z}_{K_3}]^2} + \frac{g^{*}_{\alpha,\beta,4}(n)}{E[\mathbf{Z}_{K_3}]^2} + \frac{g^{**}_{\alpha,\beta,5}(n)}{E[\mathbf{Z}_{K_3}]^2} + \frac{\binom{n}{3}\binom{n-3}{3}(P[\text{the triangle } uvw])^2 - E[\mathbf{Z}_{K_3}]^2}{E[\mathbf{Z}_{K_3}]^2}.$$

So, if each of the four terms goes to zero or is negative, then $P[\mathbf{Z}_{K_3} = 0] \to 0$.

Now, let $(\beta, \alpha) \in \zeta(f'^{K_3})$. Then $E[\mathbf{Z}_{K_3}] \to \infty$ and clearly the first term in the sum $\frac{E[\mathbf{Z}_{k_3}]}{E[\mathbf{Z}_{K_3}]^2} = \frac{1}{E[\mathbf{Z}_{k_3}]} \to 0$. Also, the last term of the sum is

$$\frac{\binom{n}{3}\binom{n-3}{3}(P[\text{the triangle } uvw])^2 - E[\mathbf{Z}_{K_3}]^2}{E[\mathbf{Z}_{K_3}]^2} =$$

$$= \frac{\binom{n}{3}\binom{n-3}{3}(P[\text{the triangle } uvw])^2 - \binom{n}{3}^2(P[\text{the triangle } uvw])^2}{E[\mathbf{Z}_{K_3}]^2} \leq 0.$$

So, if we can show that $\frac{g^*_{\alpha,\beta,4}(n)}{E[\mathbf{Z}_{K_3}]^2} \to 0$ and $\frac{g^{**}_{\alpha,\beta,5}(n)}{E[\mathbf{Z}_{K_3}]^2} \to 0$ then we will have shown that $P[\mathbf{Z}_{K_3} = 0] \to 0$.

Now, recall from the proof of Proposition 5.3.3 that $E[\mathbf{Z}_{K_3}] \asymp f^{K_3}_{\alpha,\beta,3}(n) = n^{3-3\alpha-2\beta} + n^{3-5\alpha-\beta} + n^{3-6\alpha}$ and that $f^{K_3}_{\alpha,\beta,3}(n) \equiv f'^{K_3}_{\alpha,\beta,3}(n)$. So we see that $E[\mathbf{Z}_{K_3}]^2 \asymp (f^{K_3}_{\alpha,\beta,3}(n))^2$ and, noting that all of the coefficients of α are distinct and increasing as the coefficients of β decrease, we use Proposition 4.2.6 to yield

$$E[\mathbf{Z}_{K_3}]^2 \asymp (f^{K_3}_{\alpha,\beta,3}(n))^2 \equiv f^{K_3}_{\alpha,\beta,3}(n) \equiv f'^{K_3}_{\alpha,\beta,3}(n).$$

Hence the convex set associated with $E[\mathbf{Z}_{K_3}]^2$ is the same as the set associated with $E[\mathbf{Z}_{K_3}]$, i.e., $\mathcal{C}(f'^{K_3})$.

We show that $\frac{g^*_{\alpha,\beta,4}(n)}{E[\mathbf{Z}_{K_3}]^2} \to 0$ and $\frac{g^{**}_{\alpha,\beta,5}(n)}{E[\mathbf{Z}_{K_3}]^2} \to 0$. First let us consider $\frac{g^*_{\alpha,\beta,4}(n)}{E[\mathbf{Z}_{K_3}]^2} \asymp \frac{g^*_{\alpha,\beta,4}(n)}{(f^{K_3}_{\alpha,\beta,3}(n))^2}$. By Theorem 4.3.5 $\frac{g^*_{\alpha,\beta,4}(n)}{(f^{K_3}_{\alpha,\beta,3}(n))^2} \to 0$ if all of the lines associated with the terms of $g^*_{\alpha,\beta,4}(n)$ are 'below' $\mathcal{C}((f^{K_3})^2) = \mathcal{C}(f'^{K_3})$.

It is easy to see by Figure 5.7 that this is indeed the case. Therefore

$$\frac{g^*_{\alpha,\beta,4}(n)}{E[\mathbf{Z}_{K_3}]^2} \asymp \frac{g^*_{\alpha,\beta,4}(n)}{(f^{K_3}_{\alpha,\beta,3}(n))^2} \to 0$$

for all $(\beta, \alpha) \in \zeta(f'^{K_3})$.

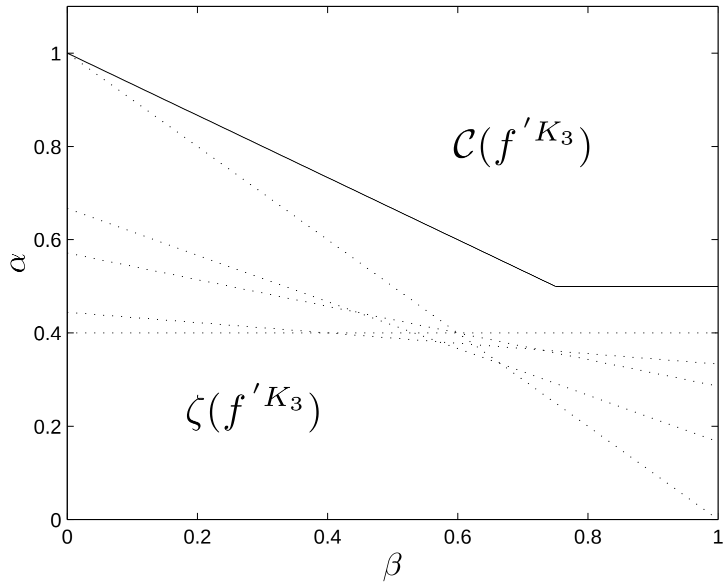


Figure 5.7: The dotted lines are those associated with terms of $g^*_{\alpha,\beta,4}(n)$ and are all 'below' $\mathcal{C}(f'^{K_3})$.

Also, by an argument similar to that above and Figure 5.8, we see that

$$\frac{g^{**}_{\alpha,\beta,5}(n)}{E[\mathbf{Z}_{K_3}]^2} \asymp \frac{g^{**}_{\alpha,\beta,5}(n)}{(f^{K_3}_{\alpha,\beta,3}(n))^2} \to 0$$

for all $(\beta, \alpha) \in \zeta(f'^{K_3})$.

So, $P[\mathbf{Z}_{K_3} = 0] \to 0$ as $n \to \infty$ whenever $(\beta, \alpha) \in \zeta(f'^{K_3})$ and almost every graph contains a K_3. Hence a threshold function for the appearance of K_3 in G is the boundary of the convex region $\mathcal{C}(f'^{K_3})$.

QED.

We would like to prove a general threshold result, however without a result for the POP of a general graph, we have as yet been unable. Instead we present a result

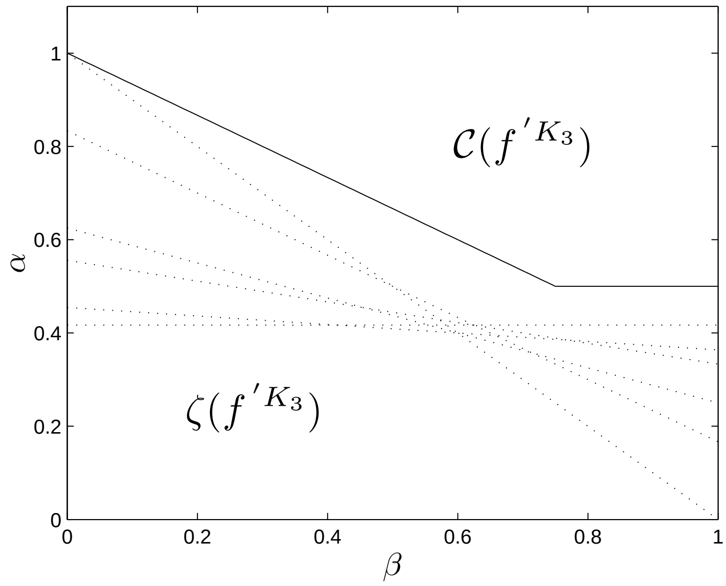


Figure 5.8: The dotted lines are those associated with terms of $g^{**}_{\alpha,\beta,5}(n)$ and are all 'below' $\mathcal{C}(f^{K_3})$.

that gives a method to prove threshold results when the threshold is the boundary of the associated convex set and all required POPs are known. But first, we prove a lemma that allows us to use Proposition 4.2.6.

Lemma 5.3.6. *Let H be a connected graph with $|E(H)| = m$ edges. Let*

$$g_H(p, 1/t) = p^{x_0} + \frac{p^{x_1}}{t} + \frac{p^{x_2}}{t^2} + \cdots + \frac{p^{x_{m-2}}}{t^{m-2}} + \frac{p^{x_{m-1}}}{t^{m-1}}$$

be the POP of H. Let $i, j \in \{0, 1, \ldots, m-1\}$. If $i < j$ then $x_i > x_j$.

Proof:

By Lemma 5.2.5 for all $0 \leq j \leq m-1$, $x_j = h_{m-j}$ is the minimum number of vertices used in a graph induced by a partition of the edge set into of $m - j$ parts.

Additionally, Lemma 5.2.10 tells us that whenever $i < j$, $x_i = h_{m-i} > h_{m-j} = x_j$. QED.

Theorem 5.3.7. *Let G be drawn from $\mathcal{D}[n, \{0,1\}^t, p]$ with $p = \frac{1}{n^\alpha}$ and $t = n^\beta$ for $\alpha, \beta \geq 0$. Let H be a connected graph with $|V(H)| = h$ vertices and $|E(H)| = m$ edges. Let*

$$g_H(p, 1/t) = p^{x_0} + \frac{p^{x_1}}{t} + \frac{p^{x_2}}{t^2} + \cdots + \frac{p^{x_{m-2}}}{t^{m-2}} + \frac{p^{x_{m-1}}}{t^{m-1}}$$

be the POP of H. Let $f^H_{\alpha,\beta,h}(n) = \sum_{i=0}^{m-1} n^{h - x_i\alpha - \beta i}$ and $\mathcal{C}(f^H)$ be the convex region in which $f^H_{\alpha,\beta,h}(n) \to 0$ as described in Proposition 4.1.1.

Let $\mathcal{H}$ be the set of all graphs that can be formed by merging two copies of H. For any $A \in \mathcal{H}$, let $h_A = |V(A)|$ and $m_A = |E(A)|$. Also, suppose that the POP of A is

$$g_A(p, 1/t) = p^{x(A)_0} + \frac{p^{x(A)_1}}{t} + \frac{p^{x(A)_2}}{t^2} + \cdots + \frac{p^{x(A)_{m_A-2}}}{t^{m_A-2}} + \frac{p^{x(A)_{m_A-1}}}{t^{m_A-1}}.$$

Let $f^A_{\alpha,\beta,h_A}(n) = \sum_{i=0}^{m_A-1} n^{h_A - x(A)_i\alpha - \beta i}$. Finally, for all $0 \leq i \leq m_A - 1$, let $l(A)_i : h_A - x(A)_i\alpha - \beta i = 0$ be the line associated with the ith term in the sum $f^A_{\alpha,\beta,h_A}(n)$.

If for all $A \in \mathcal{H}$ with $h < |V(A)| < 2h$, all of the lines $l(A)_i$, $0 \leq i \leq m_A$, are 'below' $\mathcal{C}(f^H)$ then a threshold function for the appearance of H in G is $p = \frac{1}{n^\alpha}$, where $\alpha = F^H(\beta)$ is the boundary of the convex region $\mathcal{C}(f^H)$. That is as $n \to \infty$, with high probability G will contain no copies of H whenever $(\beta, \alpha) \in \mathcal{C}(f^H)$ and G will have at least one copy of H whenever $(\beta, \alpha) \in \zeta(f^H)$.

Proof:

Let Z_H be the number of copies of H in G. Then, as discussed in section Section 5.3.1, $E[\mathbf{Z}_H] \asymp \sum_{i=0}^{m-1} n^{h-x_i\alpha-\beta i} = f^H_{\alpha,\beta,h}(n)$. Then Proposition 4.1.1 implies that $E[\mathbf{Z}_H] \to 0$ whenever $(\beta,\alpha) \in \mathcal{C}(f^H)$ and $E[\mathbf{Z}_H] \to \infty$ whenever $(\beta,\alpha) \in \zeta(f^H)$. Hence, the piecewise linear boundary function $\alpha = F^H(\beta)$ is a our candidate for the threshold of the appearance of H in G.

Now, by Markov's inequality $P[\mathbf{Z}_H \geq 1] \leq E[\mathbf{Z}_H] \to 0$ as $n \to \infty$ whenever $(\beta,\alpha) \in \mathcal{C}(f^H)$. Therefore $P[\mathbf{Z} = 0] \to 1$ and almost no G contains H whenever $(\beta,\alpha) \in \mathcal{C}(f^H)$.

Also, whenever $(\beta,\alpha) \in \zeta(f^H)$, $E[\mathbf{Z}_H] \to \infty$ as $n \to \infty$ and we use the second moment method. By Chebychev's inequality we know that

$$P[\mathbf{Z}_H = 0] \leq \frac{\mathrm{Var}(\mathbf{Z}_H)}{E[\mathbf{Z}_H]^2} = \frac{E[\mathbf{Z}_H^2] - E[\mathbf{Z}_H]^2}{E[\mathbf{Z}_H]^2}$$

and so, we calculate $E[\mathbf{Z}_H^2]$ and $E[\mathbf{Z}_H]^2$.

Now, for all $A \in \mathcal{H}$ let $\mathbf{Z}_A$ be the number of copies of A in G. Then we claim that $E[\mathbf{Z}_H^2] = E[\sum_{A\in\mathcal{H}} \mathbf{Z}_A]$.

Let $\mathcal{B}$ be the set of all graphs that lie on a subset of the vertex set of G and are isomorphic to H. Then for each $B \in \mathcal{B}$, let $\mathbf{I}_B$ be the indicator that $B \subseteq G$. Then

$\mathbf{Z}_H = \sum_{B \in \mathcal{B}} \mathbf{I}_B$ and

$$E[\mathbf{Z}_H^2] = E\left[\left(\sum_{B \in \mathcal{B}} \mathbf{I}_B\right)^2\right] = E\left[\left(\sum_{B \in \mathcal{B}} \mathbf{I}_B\right)\left(\sum_{B^* \in \mathcal{B}} \mathbf{I}_{B^*}\right)\right] = E\left[\sum_{B,B^* \in \mathcal{B}} \mathbf{I}_B \, \mathbf{I}_{B^*}\right].$$

Now, $\mathbf{I}_B \, \mathbf{I}_{B^*}$ is equal to 1 if and only if both B and B^* appear in G, i.e. if and only if $B \bigcup B^*$ appears in G. Therefore, $\mathbf{I}_B \, \mathbf{I}_{B^*} = \mathbf{I}_{B \bigcup B^*}$ where $\mathbf{I}_{B \bigcup B^*}$ is the indicator function for the graph $B \bigcup B^*$ in G.

Hence

$$E[\mathbf{Z}_H^2] = E\left[\sum_{B,B^* \in \mathcal{B}} \mathbf{I}_{B \cup B^*}\right] = E\left[\sum_{A \in \mathcal{H}} \sum_{\substack{B,B^* \in \mathcal{B} \\ B \bigcup B^* = A}} \mathbf{I}_{B \cup B^*}\right] = \sum_{A \in \mathcal{H}} E\left[\sum_{\substack{B,B^* \in \mathcal{B} \\ B \bigcup B^* = A}} \mathbf{I}_{B \cup B^*}\right]$$

$$= \sum_{A \in \mathcal{H}} \sum_{\substack{B,B^* \in \mathcal{B} \\ B \bigcup B^* = A}} E[\mathbf{I}_{B \cup B^*}] = \sum_{A \in \mathcal{H}} \sum_{\substack{B,B^* \in \mathcal{B} \\ B \bigcup B^* = A}} P_{\geq}[B \cup B^*] = \sum_{A \in \mathcal{H}} \sum_{\substack{B,B^* \in \mathcal{B} \\ B \bigcup B^* = A}} P_{\geq}[A]$$

$$= \sum_{A \in \mathcal{H}} \sum_{\substack{B,B^* \in \mathcal{B} \\ B \bigcup B^* = A}} E[\mathbf{I}_A] = \sum_{A \in \mathcal{H}} E\left[\sum_{\substack{B,B^* \in \mathcal{B} \\ B \bigcup B^* = A}} \mathbf{I}_A\right] = \sum_{A \in \mathcal{H}} E[\mathbf{Z}_A]$$

and our claim is true. Thus, $E[\mathbf{Z}_H^2] = E[\sum_{A \in \mathcal{H}} \mathbf{Z}_A]$.

Now,

$$E[\mathbf{Z}_H^2] = \sum_{A \in \mathcal{H}} E[\mathbf{Z}_A] = \sum_{\substack{A \in \mathcal{H} \\ |V(A)| = h}} E[\mathbf{Z}_A] + \sum_{\substack{A \in \mathcal{H} \\ h+1 \leq |V(A)| = 2h-1}} E[\mathbf{Z}_A] + \sum_{\substack{A \in \mathcal{H} \\ |V(A)| = 2h}} E[\mathbf{Z}_A].$$

If $|V(A)| = h$, then $A = H$ and so $\sum_{\substack{A \in \mathcal{H} \\ |V(A)| = h}} E[\mathbf{Z}_A] = E[\mathbf{Z}_H]$. Also, if $|V(A)| = 2h$

then A is the graph that contains two disjoint copies of H and so $\sum_{\substack{A\in\mathcal{H}\\|V(A)|=2h}} E[\mathbf{Z}_A] = \binom{n}{h}\binom{n-h}{h}P_{\geq}[H]^2$. So, we have that

$$E[\mathbf{Z}_H^2] = E[\mathbf{Z}_H] + \sum_{\substack{A\in\mathcal{H}\\h+1\leq|V(A)|=2h-1}} E[\mathbf{Z}_A] + \binom{n}{h}\binom{n-h}{h}P_{\geq}[H]^2.$$

Therefore Chebychev's inequality becomes

$$P[\mathbf{Z}_H = 0] \leq \frac{E[\mathbf{Z}_H] + \sum_{\substack{A\in\mathcal{H}\\h+1\leq|V(A)|=2h-1}} E[\mathbf{Z}_A] + \binom{n}{h}\binom{n-h}{h}P_{\geq}[H]^2 - E[\mathbf{Z}_H]^2}{E[\mathbf{Z}_H]^2}$$

$$= \frac{E[\mathbf{Z}_H]}{E[\mathbf{Z}_H]^2} + \frac{\sum_{\substack{A\in\mathcal{H}\\h+1\leq|V(A)|=2h-1}} E[\mathbf{Z}_A]}{E[\mathbf{Z}_H]^2} + \frac{\binom{n}{h}\binom{n-h}{h}P_{\geq}[H]^2 - E[\mathbf{Z}_H]^2}{E[\mathbf{Z}_H]^2}.$$

Now, let us consider each of the three terms separately. The first term $\frac{E[\mathbf{Z}_H]}{E[\mathbf{Z}_H]^2} \to 0$ whenever $(\beta,\alpha) \in \zeta C(f^H)$ since $E[\mathbf{Z}_H] \to \infty$. Also, the last term is

$$\frac{\binom{n}{h}\binom{n-h}{h}P_{\geq}[H]^2 - E[\mathbf{Z}_H]^2}{E[\mathbf{Z}_H]^2} = \frac{\binom{n}{h}\binom{n-h}{h}P_{\geq}[H]^2 - \binom{n}{h}^2 P_{\geq}[H]^2}{E[\mathbf{Z}_H]^2} \leq 0.$$

Therefore, if we can show that the middle term $\left(\sum_{\substack{A\in\mathcal{H}\\h+1\leq|V(A)|=2h-1}} E[\mathbf{Z}_A]\right)/E[\mathbf{Z}_H]^2 \to 0$ whenever $(\beta,\alpha) \in \mathcal{C}(f^H)$, then $P[\mathbf{Z}_H = 0] \to 0$ and we will be done.

Now, there are only a finite number of $A \in \mathcal{H}$. Therefore if we can show that for all $A \in \mathcal{H}$ with $h+1 \leq |V(A)| = 2h-1$,we have $E[\mathbf{Z}_A]/E[\mathbf{Z}_H]^2 \to 0$ then $\left(\sum_{\substack{A\in\mathcal{H}\\h+1\leq|V(A)|=2h-1}} E[\mathbf{Z}_A]\right)/E[\mathbf{Z}_H]^2 \to 0$.

So, let $A \in \mathcal{H}$ with $h+1 \leq |V(A)| = 2h-1$, then as discussed in Section 5.3.1

$$E[\mathbf{Z}_A] \asymp \binom{n}{h_A} g_A(p, 1/t) \asymp \sum_{i=0}^{m_A-1} n^{h_A - x(A)_i\alpha - \beta i} = f^A_{\alpha,\beta,h_A}(n)$$

and so $E[\mathbf{Z}_A]/E[\mathbf{Z}_H]^2 \to 0$ if and only if $f^A_{\alpha,\beta,h_A}(n)/E[\mathbf{Z}_H]^2 \to 0$.

Now, $E[\mathbf{Z}_H] \asymp f^H_{\alpha,\beta,h}(n) = \sum_{i=0}^{m-1} n^{h-x_i\alpha-\beta i}$ and so

$$E[\mathbf{Z}_H]^2 \asymp (f^H_{\alpha,\beta,h}(n))^2 = \left(\sum_{i=0}^{m-1} n^{h-x_i\alpha-\beta i}\right)^2 = \sum_{i,j=0}^{m-1} n^{2h-(x_i+x_j)\alpha-\beta(i+j)}$$

and is a posynomial of the form discussed in Chapter 4.

Furthermore, whenever $(\beta, \alpha) \in \zeta(f^H)$, $(f^H_{\alpha,\beta,h}(n))^2 \to \infty$ and

$$\frac{E[\mathbf{Z}_A]}{E[\mathbf{Z}_H]^2} \asymp \frac{f^A_{\alpha,\beta,h_A}(n)}{(f^H_{\alpha,\beta,h}(n))^2} = \frac{\sum_{i=0}^{m_A-1} n^{h_A-x(A)_i\alpha-\beta i}}{\sum_{i,j=0}^{m-1} n^{2h-(x_i+x_j)\alpha-\beta(i+j)}}$$

is a ratio of the form[2] discussed in Theorem 4.3.5 since $h_A < 2h$. So, $\frac{E[\mathbf{Z}_A]}{E[\mathbf{Z}_H]^2} \to 0$ if for each $0 \leq i \leq m_A$, l_i is 'below' $\mathcal{C}((f^H)^2)$.

Finally, by Lemma 5.3.6, we have that $x_i > x_j$ when $i < j$, and so by Proposition 4.2.6 we have that $(f^H_{\alpha,\beta,h}(n))^2 \equiv f^H_{\alpha,\beta,h}(n)$ and so $\mathcal{C}((f^H)^2) = \mathcal{C}(f^H)$. Therefore $\frac{E[\mathbf{Z}_A]}{E[\mathbf{Z}_H]^2} \to 0$ if for each $0 \leq i \leq m_A$, l_i is 'below' $\mathcal{C}(f^H)$.

Hence, if for all $A \in \mathcal{H}$ with $h < |V(A)| < 2h$ all of the lines $l(A)_i$, $0 \leq i \leq m_A$, are 'below' $\mathcal{C}(f^H)$ then $P[\mathbf{Z}_H = 0] \to 0$ as $n \to \infty$ whenever $(\beta, \alpha) \in \zeta(f^H)$ and

[2] While it is true that the upper limits on the sums do not agree, this can easily be fixed by setting both upper limits to some M that is larger than both $(m-1)^2$ and m_A and then adding terms equal to 0 to fill out the summation.

almost every G will contain a copy of H.

Therefore, if for all $A \in \mathcal{H}$ with $h < |V(A)| < 2h$, all of the lines $l(A)_i$, $0 \leq i \leq m_A$, are 'below' $\mathcal{C}(f^H)$, then, as $n \to \infty$, with high probability G will be contain no copies of H whenever $(\beta, \alpha) \in \mathcal{C}(f^H)$ and G will have at least one copy of H whenever $(\beta, \alpha) \in \zeta(f^H)$. So a threshold function for the appearance of H in G is $\alpha = F^H(\beta)$, the boundary of the convex region $\mathcal{C}(f^H)$.

QED.

Theorem 5.3.7 gives us a method to prove threshold results if the boundary of the convex region associated with the graph is indeed the threshold and all needed POPs are known. This might cause one to conjecture that the boundary is always the threshold, but this is not the case.

Consider the kite graph H_{kite} discussed in Section 5.3.1. By Proposition 5.3.4 we know that the boundary function for $\mathcal{C}(f'^{H_{kite}}) = \mathcal{C}(f^{H_{kite}})$ is

$$\alpha = \begin{cases} \frac{5-6\beta}{5} & \beta \leq \frac{5}{11} \\ \frac{5-5\beta}{6} & \frac{5}{11} \leq \beta \leq \frac{4}{7} \\ \frac{5}{14} & \beta \geq \frac{4}{7} \end{cases}$$

and $E[\mathbf{Z}_{kite}] \to 0$ whenever $(\beta, \alpha) \in \mathcal{C}(f'^{H_{kite}})$ and $E[\mathbf{Z}_{kite}] \to \infty$ whenever $(\beta, \alpha) \in \zeta(f'^{H_{kite}})$. However $p = \frac{1}{n^\alpha}$ is not the correct threshold.

First we show that H_{kite} does not satisfy all of the conditions of Theorem 5.3.7. Consider the graph H_{2-kite} with vertex set $V(H_{2-kite}) = \{v_1, v_2, v_3, v_4, v_5, v_6\}$ and edge set

$$E(H_{2-kite}) = \{\{v_1, v_2\}, \{v_1, v_3\}, \{v_1, v_4\}, \{v_1, v_5\}, \{v_2, v_3\}, \{v_2, v_4\}, \{v_3, v_4\}, \{v_3, v_6\}\}.$$

H_{2-kite} is a K_4 with two pendant edges attached to different vertices, see Figure 5.9. Also, H_{2-kite} is a graph that can be formed by merging to copies of H_{kite}, i.e., $H_{2-kite} \in \mathcal{H}_{kite}$.

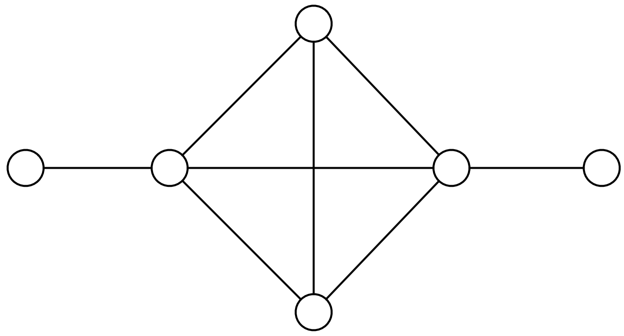


Figure 5.9: The graph H_{2-kite}.

It is easy to show, either directly using Lemma 5.2.5 or by applying the pendant edge Lemma 5.2.7 to the POP of H_{kite}, that the POP of H_{2-kite} is $g_{H_{2-kite}}(p, 1/t) = p^{16} + \frac{p^{15}}{t} + \frac{p^{13}}{t^2} + \frac{p^{12}}{t^3} + \frac{p^{10}}{t^4} + \frac{p^8}{t^5} + \frac{p^7}{t^6} + \frac{p^6}{t^7}$. Therefore, the expected number of copies of H_{kite} in G is

$$E[\mathbf{Z}_{H_{2-kite}}] \asymp \binom{n}{6} g_{H_{2-kite}}(p, 1/t) \asymp f^{H_{2-kite}}_{\alpha,\beta,6}(n)$$

$$= n^{6-16\alpha} + n^{6-15\alpha-\beta} + n^{6-13\alpha-2\beta} + n^{6-12\alpha-3\beta} + n^{6-10\alpha-4\beta} + n^{6-8\alpha-5\beta} + n^{6-7\alpha-6\beta} + n^{6-6\alpha-7\beta}.$$

Now, consider the line $l(H_{2-kite})_0 : 6 - 16\alpha = 0$ associated with the first term in the POP of H_{2-kite}. In Figure 5.10 we see that $l(H_{2-kite})_0$ does not always fall 'below' the convex set $\mathcal{C}(f^{H_{kite}})$ and hence Theorem 5.3.7 does not apply.

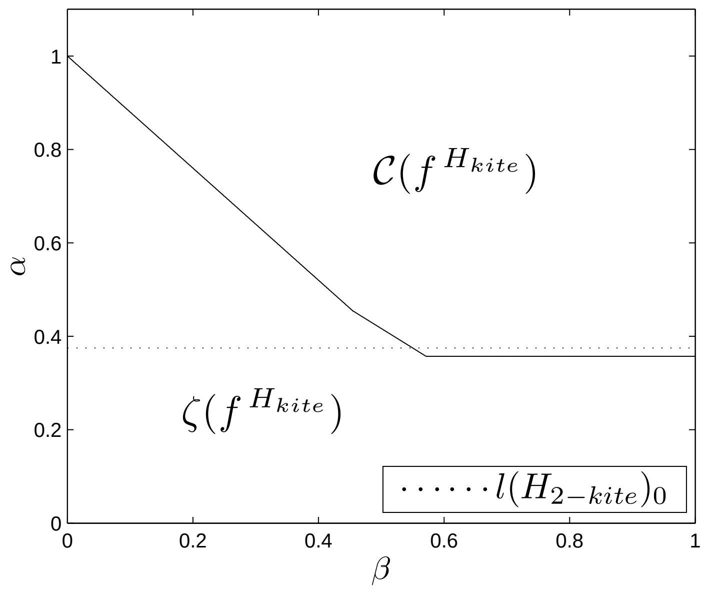


Figure 5.10: The line $l(H_{2-kite})_0$ is not always 'below' $\mathcal{C}(f^{H_{kite}})$.

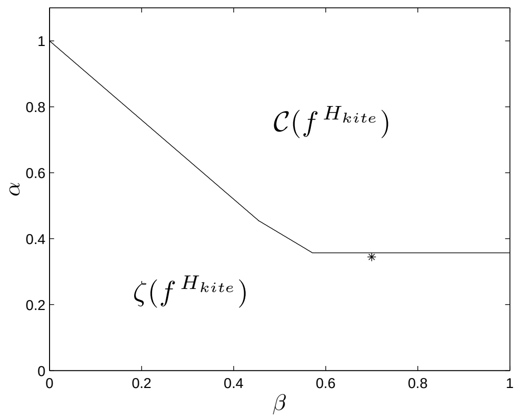
Additionally H_{kite} cannot appear in G if K_4 does not appear. Now, by Proposition 5.3.3 we know that

$$E[\mathbf{Z}_{K_4}] \equiv f'^{K_4}_{\alpha,\beta,4}(n) = n^{4-4\alpha-5\beta} + n^{4-12\alpha}$$

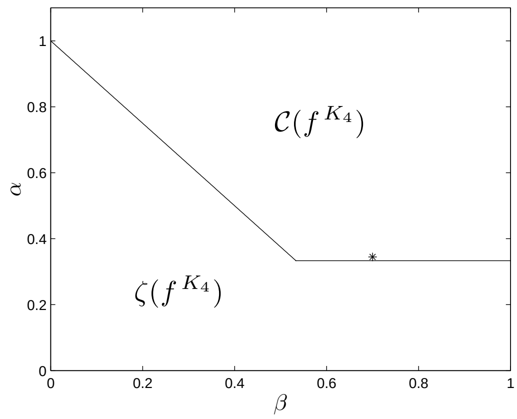
and $E[\mathbf{Z}_{K_4}] \to 0$ whenever $(\beta, \alpha) \in \mathcal{C}(f'^{K_4})$.

Now, consider $(\beta, \alpha) = (7/10, 10/29)$. $(7/10, 10/29) \in \zeta(f^{H_{kite}})$ and so $E[\mathbf{Z}_{H_{kite}}] \to \infty$ as $n \to \infty$. If $\alpha = F'^{H_{kite}}(\beta)$, the boundary of the convex region $\mathcal{C}(f'^{H_{kite}})$ was the correct threshold, then at $(7/10, 10/29)$ almost every G would contain a copy of

H_{kite}. However, $(7/10, 10/29) \in \mathcal{C}(f^{K_4})$and so $E[\mathbf{Z}_{K_4}] \to 0$ as $n \to \infty$. Therefore, by Markov's inequality, $P[\mathbf{Z}_{k_4} \geq 1] \leq E[\mathbf{Z}_{k_4}] \to 0$ and almost no G contains a K_4. Hence, almost no G contains a H_{kite}. Therefore, the boundary of the convex region is not the correct threshold.



(a)



(b)

Figure 5.11: **(a)** The point $(7/10, 10/29) \in \mathcal{C}(f^{H_{kite}})$. **(b)** The point $(7/10, 10/29) \in \zeta(f^{K_4})$.

Although we have been unable to prove a more general threshold result, we make the following conjectures.

Conjecture 5.3.8. *Let G be drawn from $\mathcal{D}[n, \{0,1\}^t, p]$ with $p = \frac{1}{n^\alpha}$ and $t = n^\beta$ for $\alpha, \beta \geq 0$. Let H be a connected graph and $\mathcal{H}^*$ be the set of all non empty subgraphs of H. Let Λ_H be the region in the β, α-parameter space in which for all $H^* \in \mathcal{H}^*$, $E[\mathbf{Z}_{H^*}] \to \infty$ as $n \to \infty$. Then the threshold for the appearance of H is the boundary of Λ_H.*

Recall from Section 3.2.1, for any graph H on k vertices and $l \geq 1$ edges, $\varepsilon(H) = \frac{l}{k}$.

We say that H is *strictly balanced* if for every proper subgraph H' of H, $\varepsilon(H') < \varepsilon(H)$. We make the following conjecture about the threshold for the appearance of a strictly balanced graph H as a subgraph of a Discrete Random Dot Product Graph.

Conjecture 5.3.9. *Let G be drawn from $\mathcal{D}[n, \{0,1\}^t, p]$ with $p = \frac{1}{n^\alpha}$ and $t = n^\beta$ for $\alpha, \beta \geq 0$. Let H be a connected graph with $|V(H)| = h$ vertices and $|E(H)| = m$ edges. Let*

$$g_H(p, 1/t) = p^{x_0} + \frac{p^{x_1}}{t} + \frac{p^{x_2}}{t^2} + \cdots + \frac{p^{x_{m-2}}}{t^{m-2}} + \frac{p^{x_{m-1}}}{t^{m-1}}$$

be the POP of H. Let $f^H_{\alpha,\beta,h}(n) = \sum_{i=0}^{m-1} n^{h - x_i\alpha - \beta i}$ and $\alpha = F^H(\beta)$ be the boundary of the convex region in which $f^H_{\alpha,\beta,h}(n) \to 0$ as described in Proposition 4.1.1. If H is strictly balanced, then a threshold function for the appearance of H in G is $p = \frac{1}{n^\alpha}$.

Chapter 6

Overview and Directions for Future Work

In this thesis we introduce and examine the Random Dot Product Graph suite of models. Three sets of models for social networks are presented, all based on the idea of common or shared interest. The first model, discussed in Chapter 2, is the Dense Random Dot Product Graph. In this model, all small subgraphs appear with probability approaching one as the number of vertices approaches infinity. The second model, discussed in Chapter 3, is the Sparse Random Dot Product Graph and small subgraphs appear at certain thresholds. Finally, the third model is a discrete version and is discussed in Chapter 5. As in the sparse model, in the Discrete Random Dot Product Graph small subgraphs appear at various thresholds.

In this chapter, we revisit each of the models, review our results and suggest ideas for further work. Then in the last section we discuss more general ideas and directions.

6.1 Observations and New Directions

In this section, we recap the main results and pose open questions related to each of the individual models.

6.1.1 The Dense Random Dot Product Graph

In Chapter 2 we introduce the general Random Dot Product Graph. We then focus on the Dense model for the dimension one case, i.e. when G is drawn from $\mathcal{D}[n,a,1]$. In Section 2.2.1, we show that the expected number of edges is $\binom{n}{2}\frac{1}{(a+1)^2}$ and for a fixed k, the expected number of vertices of degree k, denoted $\lambda(k)$, is $E[\lambda(k)] \sim \left(\frac{1}{k!\,a}(1+a)^{\frac{1}{a}}\Gamma\left(\frac{1}{a}+k\right)\right)n^{\frac{a-1}{a}}$ as $n \to \infty$. However, we do not present the variance calculations, the next logical step, and so we pose the questions:

- How close to expected are these values?
- If we allow $0 < a \leq 1$, then for what k do we, with high probability, have vertices of degree k, i.e., what is the threshold (in terms of a) for the appearance of vertices of degree k?

In Section 2.3.1, we show that the model obeys the degree distribution power law. However, we do not directly prove Conjecture 2.3.1 and hence it is left for future work.

In Sections 2.3.2 and 2.3.3 we show that the Dense Random Dot Product Graph exhibits clustering and has a low fixed diameter of at most six as $n \to \infty$. Also, recall, by the definition presented in Section 1.2.4, that a small world graph is a graph G on n vertices with average degree k for which $L_G \approx L_{R(n,k)} \approx \frac{\log(n)}{\log(k)}$, but $\gamma_G \gg \gamma_{R(n,k)} \approx \frac{k}{n}$, that is a graph for which the characteristic path length is small and similar to that of a random graph, but whose vertices are highly clustered.

For the Dense Random Dot Product Graph drawn from $\mathcal{D}[n, a, 1]$ the expected average degree is $E[k] = (n-1)/(a+1)^2$. So, by small world standards, G would be highly clustered if it's clustering coefficient $\gamma_G \gg \frac{n-1}{n(a+1)^2}$ and we posit the following:

Conjecture 6.1.1. *Let G be drawn from $\mathcal{D}[n, a, 1]$. Then $P(\gamma_G \gg \frac{1}{(a+1)^2})$ tends to* 1 *as $n \to \infty$.*

Empirical results support this conjecture, as does our clustering result Lemma 2.3.6. The other main characteristic of a small world graph is that the characteristic path length L_G is approximately that of the random graph $R(n, k)$, where $L_R(n, k) \approx \frac{\log(n)}{\log\left(\frac{n+1}{(a+1)^2}\right)} \approx 1$ for large n. Ignoring vertices of degree zero, our diameter result, Theorem 2.3.9, has shown that for G drawn from $\mathcal{D}[n, a, 1]$ one has $L_G = \theta(1)$.

Towards the end of Chapter 2, we extend the Dense Random Dot Product Graph into higher dimensions, i.e., consider the case when G is drawn from $\mathcal{D}[n, a, t]$, $(t > 0)$. We prove some results similar to those in a single dimension. We again show that the expected number of isolated vertices is $\binom{n}{2}/(a+1)^2$ and is not dependent on the dimension t. Also, we show that as $n \to \infty$ and for a fixed k, the expected number of vertices of degree k is $E[\lambda(k)] \sim C(k,t,a)n^{\frac{a-t}{a}}$ where $C(k,t,a)$ is a constant depending only on k, t, and a. However, as in the one dimensional case, we do not discuss the variance calculations and so the same questions with regards to actual values and the thresholds for the appearance of such vertices persist.

Finally, we end the chapter by discussing a bend that occurs in the degree distribution power law whenever $t > 1$. While we discuss a possible reason as to why this bend occurs, we did not prove Conjecture 2.4.3, and so it is left as future work.

6.1.2 The Sparse Random Dot Product Graph

In Chapter 3 we introduce and examine the Sparse Random Dot Product Graph where G is drawn from $\mathcal{DS}[n,a,b,1]$ and b is restricted to some subset of $(0,\infty)$. We begin the chapter by discussing some general results for any $b \in (0,\infty)$. In Section 3.1.1 we show that the expected number of edges is $\binom{n}{2}/(n^b\,(a+1)^2)$ and that a threshold for the appearance of edges is $b = 2$. We also show that a threshold for the appearance of a clique of size k is $\frac{b}{k-1}$ and hence the clique number of G is $\omega(G) \in \{k_b, k_b + 1\}$ where $k_b = \max\{k : b < \frac{2}{k-1}, k \in \mathbb{Z}_+\}$.

In Section 3.1.2 we restrict b to the interval $(0,1)$. We show that, for a fixed k, the expected number of vertices of degree k is $E[\lambda(k)] \sim \left(\frac{1}{k!\,a}(1+a)^{\frac{1}{a}}\Gamma\left(\frac{1}{a}+k\right)\right)n^{\frac{a-1+b}{a}}$ as $n \to \infty$. As in the dense model, we do not discuss the variance calculations for a general k, and so we again pose the same questions with regards to the true value of $\lambda(k)$ and thresholds. However, we do present a variance result for the case $k = 0$, when we are concerned with the number of isolated vertices. We show that with probability tending to one as $n \to \infty$, G always has isolated vertices and hence is never connected. We also show that the sparse model exhibits clustering.

In Section 3.1.3 we begin by presenting a general result relating the number of edges in a fixed graph to the probability of the graph appearing as a subgraph of G. Specifically, we show that for a graph H, the probability of H appearing as a (not necessarily induced) subgraph of a Sparse Random Dot Product Graph G is $\frac{C(H)}{n^{mb}}$ where m is the number of edges in H and $C(H)$ is a constant depending only on the graph H. Next we show that a threshold for the appearance of a tree on k vertices is $b = \frac{k}{k-1}$. Finally, we restrict b to the interval $(1, 2)$ and show that with high probability G contains no cycles and therefore is a forest on trees of size $\frac{b}{b-1}$ or less.

In Section 3.2 we present the main results for the Sparse Random Dot Product Graph. In Section 3.2.1 we show that for a graph H on k vertices and $l \geq 1$ edges, a threshold for the appearance of H is $b = \frac{1}{\varepsilon'(H)}$. In Section 3.2.2, we show the Sparse Random Dot Product Graph obeys the degree distribution power law whenever $b \in (0, 1)$. Additionally, we know that, with high probability, G is edgeless when $b > 2$ and so we pose the question:

- Can we describe the degree distribution of G when $b \in [1, 2]$?

In Section 3.2.3 we show that when $0 < b < \frac{2}{a+2}$, the Sparse Random Dot Product

Graph contains isolated vertices with the remaining vertices connected in a giant component of diameter at most $d_b = \lfloor \frac{1}{1-2\epsilon-b} \rfloor + 5$. This leads us to ask the following questions for $b \in [\frac{2}{a+2}, 2]$:

- If we ignore isolated vertices, how many components will G contain?
- Can we characterize the distribution on the orders of the components of G?
- Are we guaranteed isolated edges?
- What is the average path length between two vertices in a component of G?

Finally, in Section 3.3 we recap the sparse model and present the following evolution for the appearance of subgraphs of order k as b goes from zero to infinity:

- beginning at $b = 0$ all subgraphs on k vertices are present,
- at $b = \frac{2}{k-1}$ cliques of size k disappear,
- at $b = 1$ k-cycles disappear,
- at $b = \frac{k}{k-1}$ trees on k vertices disappear,
- for $b > \frac{k}{k-1}$ there are no connected subgraphs on k vertices, and
- for $b > 2$ there are no edges in the graph.

Another area of exploration for the Sparse Random Dot Product Graph, is to extend the model to higher dimensions. A first step would be to mimic the extension in the Dense case.

6.1.3 The Discrete Random Dot Product Graph

We begin Chapter 5 by introducing the Discrete Random Dot Product Graph, when G is drawn from $\mathcal{D}[n, \{0,1\}^t, p]$, and proving a few basic results. We show that the probability of an edge is p^2 and that threshold for the appearance of edges is $p = 1/n$ and is independent of the dimension t from which the vectors are drawn. We also show that whenever $p < 1$, G exhibits clustering.

In Section 5.1.2 we use K_3 as an example to demonstrate the difficulty of calculations in the discrete model. We determine the probability of K_3 appearing as a subgraph and give the equation for the variance of the number of triangles, illustrating the dependence on both p and the dimension t.

In Section 5.2 we introduce the probability order polynomial (POP) of a graph, $g_H(p, 1/t)$, as a function of p and $1/t$ that is asymptotic to $P_{\geq}[H]$, the probability of H appearing as a (not necessarily induced) subgraph of a Discrete Random Dot Product Graph G. We show that for a graph H on m edges, $P_{\geq}[H] \asymp g_H(p, 1/t) = p^{x_0} + \frac{p^{x_1}}{t} + \frac{p^{x_2}}{t^2} + \cdots + \frac{p^{x_{m-2}}}{t^{m-2}} + \frac{p^{x_{m-1}}}{t^{m-1}}$ where for all $0 \leq j \leq m-1$, $x_j = h_{m-j}$ is the minimum number of vertices used in a graph induced by a partition of the edge set into $m - j$ parts. We then give formulas to calculate the POP of trees, cycles, and complete graphs. However we are, as of yet, unable to give an effective formula for a general graph H and so this is left as future work.

In an effort to obtain such a formula for a general graph, a first step could be to develop an algorithm to partition the edge set of a graph H into parts that induce graphs that use the minimum number of vertices so that Lemma 5.2.5 can be applied. The obvious algorithms such a greedy approach do not produce the desired partition in all cases. So, we pose the question:

- Can we develop an efficient algorithm to partition the edge set of a graph into parts that satisfy the conditions of Lemma 5.2.5?

The results in Section 5.3 rely heavily on the framework built in Chapter 4. We present first moment results for trees, cycles, and complete graphs. The methods used in these proofs can easily be expanded to any graph for which the POP is known. In our examples, the boundary of the region $\mathcal{C}$ is always determined by the lines associated with the first and last terms of the POP. However, we present a counterexample to this always being the case. So, we asked the question:

- When is the boundary of $\mathcal{C}$ determined by only the first and last terms of the POP?

This will only happen if the line associated with one of the middle terms is 'above' the intersection point of the lines associated with the first and last terms. So, we offer the following conjecture.

Conjecture 6.1.2. *Let G be drawn from $\mathcal{D}[n, \{0,1\}^t, p]$ with $p = \frac{1}{n^\alpha}$ and $t = n^\beta$ for $\alpha, \beta \geq 0$. Let H be a connected graph on h vertices and m edges with POP*

$$g_H(p, 1/t) = p^{x_0} + \frac{p^{x_1}}{t} + \frac{p^{x_2}}{t^2} + \cdots + \frac{p^{x_{m-2}}}{t^{m-2}} + \frac{p^{x_{m-1}}}{t^{m-1}}.$$

Let l_0 and l_{m-1} be the lines associated with the first and last terms, respectively, of the POP of H. If for all $i \in \{1, \ldots, m-2\}$, $\frac{2mh-hx_i}{2mi} \leq \frac{2m-h}{m-1}$, then the boundary of $\mathcal{C}$ will be determined by only l_0 and l_{m-1} and hence be

$$\alpha = \begin{cases} \frac{h-(m-1)\beta}{h} & \beta \leq \frac{2mh-h^2}{2m(m-1)} \\ \\ \frac{h}{2m} & \beta \geq \frac{2mh-h^2}{2m(m-1)} \end{cases}$$

where $\frac{2mh-h^2}{2m(m-1)}$ is the β coordinate of the intersection point of l_0 and l_{m-1}.

In Section 5.3.2, we present a threshold result for K_3 and in Theorem 5.3.7 describe a general method for proving threshold results if all required POPs are known. We also present an example of a graph that does not meet the requirements of Theorem 5.3.7 and so there is more work to be done. Finally, in Conjecture 5.3.8, we state that if H is a connected graph and $\mathcal{H}^*$ is the set of all non empty subgraphs of H. Then we believe that Λ_H, the region in the β, α-parameter space in which for all $H^* \in \mathcal{H}^*$, $E[\mathbf{Z}_{H^*}] \to \infty$ as $n \to \infty$ is a threshold for the appearance of H.

Although in Chapter 5 we present a good foundation of theory, many questions are still left unanswered for the Discrete Random Dot Product Graph. In addition to new questions similar to those that arise in the other two versions of the model, there are results that were proven for those versions that remain open for the discrete case. For example:

- What is the degree distribution?

- When are we guaranteed isolated vertices?
- Is the graph connected and if so, what is the diameter?
- What is the evolution of the appearance of various subgraph?

6.2 Other Open Problems and Thoughts for Future Work

In this section we present more open questions that are not necessarily specific to a single version of the model.

6.2.1 General Graph Theory Questions

There are many graph theoretic questions that remain unanswered. For example in each of the versions of the model:

- If we ignore isolated vertices, what is the connectivity of the remaining components?
- What are the graph theoretic invariants, such as clique number, independence number, domination number, chromatic number, maximum degree, etc.?

6.2.2 Variations on the Model

So far in the continuous versions of the model, we have only discussed the basic Random Dot Product Graph in which the probability mapping f has been the identity mapping, $f(x) = x$ in the dense case, or $f(x) = \frac{x}{n^b}$ in the sparse case, and the vectors are drawn from $\mathcal{U}^a[0,1]$, powers of the uniform distribution on $[0,1]$. The model can be further explored by drawing the vectors from a variety of other distributions or by changing the probability mapping. The distribution $\mathcal{U}^a[0,1]$ is the same as the Beta distribution $B(\frac{1}{a},1)$. Therefore a natural next step is to draw the vectors from other Beta distributions. Also, there are the only limitations placed on the probability mapping is that $f : \mathbb{R} \to [0,1]$, and a variety of other functions meet this criteria, e.g. $f(x) = \frac{1}{\pi}\tan^{-1} x + \frac{1}{2}$, $f(x) = \frac{1}{2}\tanh x + \frac{1}{2}$, and $f(x) = 1 - e^{-x^2}$. However, one should note that the nature of these suggested functions may make the work intractable.

Similarly, in the discrete case, changing the probability mapping, perhaps to $f(x) = \frac{x}{n^b t}$, will change the model and open up new avenues of research.

Variations can also be made upon the way in which we extend the dense model into higher dimensions. The current method allows the vectors to be drawn from the first orthant portion of a box centered at the origin. This model does not include all vector values x and y, whose dot product $x \cdot y \leq 1$. Perhaps another way of extending the model would be to let $f(x \cdot y) = |x \cdot y|$ or $f(x \cdot y) = (x \cdot y)^2$ and then study vectors drawn from an origin centered ball of radius 1 instead. Or, more simply and keeping with the original box idea, we can draw each vector x from $[\mathcal{U}^a[0,r]]^t$ instead of just

$[\mathcal{U}^a[0, \frac{1}{\sqrt[2a]{t}}]]^t$. We require that for any x, y drawn from our distribution $x \cdot y \in [0,1]$ which is guaranteed if $tr^{2a} \leq 1$. Then for any two vertices u and v with corresponding vectors $\mathbf{x}$ and $\mathbf{y}$

$$P[x \sim y] = \frac{1}{r^{2t}} \int_0^r \cdots \int_0^r (x_1^a y_1^a + \cdots + x_t^a y_t^a) dx_1 \cdots dx_t dy_1 \cdots dy_t = \frac{tr^{2a}}{(1+a)^2}$$

and unlike before, the expected number of edges $\frac{\binom{n}{2} tr^{2a}}{(1+a)^2}$ is dependent on the dimension from which we draw the vectors.

6.2.3 In Conclusion

In this thesis, we have examined the Random Dot Product Graph, a new suite of social network models. We have shown that all three versions of the model exhibit clustering, i.e., two vertices are more likely to be adjacent if they have a common neighbor, an idea found in social network theory. We have shown that in the continuous versions of the model (both the dense and sparse cases) the Random Dot Product Graph satisfies the degree distribution power law that is found in some social networks such as the internet. Also, for the continuous versions, we have shown that they have low diameter and exhibit characteristics of small world networks. Additionally, all three versions of the model differ from Erdős-Rényi random graphs by exhibiting clustering or by the constant presence of isolated vertices. Finally, we have shown that the Random Dot Product Graph is an interesting random graph model that has many avenues that have not yet been explored.

In conclusion, the Random Dot Product Graphs is a new and diverse random graph model that exhibits characteristics of a social network and is mathematically rich enough to pose interesting problems for the future.

Bibliography

[1] Lada A. Adamic. The small world web. In *Proc. 3rd European Conf. Research and Advanced Technology for Digital Libraries*, pages 443–452, 1999.

[2] William Aiello, Fan Chung, and Linyuan Lu. A random graph model for power law graphs. *Experimental Mathematics*, 10:53–66, 2001.

[3] Reka Albert, Hawoong Jeong, and Albert-Laszlo Barabàsi. Diameter of the world-wide web. *Nature*, 401:130–131, 1999.

[4] Albert-Laszlo Barabàsi and Reka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[5] Albert-Laszlo Barabàsi, Reka Albert, and Hawoong Jeong. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A*, 281:69–77, 2000.

[6] Tian Bu and Don Towsley. On distinguishing between internet power law topology generators. In *Proceedings of INFOCOM*, 2002.

[7] Ken Calvert, Matthew Doar, and Ellen Zegura. Modeling internet topology. *IEEE Transactions of Communications*, 35(6):160–163, June 1997.

[8] Ken Calvert, Ramesh Madhavan, and Ellen Zegura. A comparison of two practical multicast routing schemes. Technical Report GIT-CC-94/25, Georgia Institute of Technology, February 1994.

[9] Matthew Doar and Ian Leslie. How bad is naive multicast routing? In *Proceedings of IEEE INFOCOM '93*, pages 82–89, 1993.

[10] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. *Computer Communication Review*, 29(4):251–262, 1999.

[11] Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

[12] Bernardo Huberman and Lada Adamic. Growth dynamics of the world-wide web. *Nature*, 401:131, 1999.

[13] Cheng Jin, Qian Chen, and Sugih Jamin. Inet: Ineternet topology generator. Technical Report CSE-TR-433-00, University of Michigan, 2000.

[14] Victor Klee and David Larman. Diameters of random graphs. *Canadian Journal of Math*, 33:618–640, 1981.

[15] Jon Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. The web as a graph: measurements models and methods. In *Proceedings of the International Conference on Combinatorics and Computing*, pages 26–28, July 1999.

[16] Alberto Medina, Ibrahim Matta, and John Byers. On the origin of power laws in internet topologies. *ACM SigCOMM Computer Communication Review*, 30(2):18–28, 2000.

[17] Stanley Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.

[18] Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. Manuscript, 2001.

[19] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[20] Karen B. Singer. *Random Intersection Graphs.* PhD thesis, The Johns Hopkins University, Baltimore, Maryland, 1995.

[21] Hongsuda Tangmunarunkit, Ramesh Govindan, Sugih Jamin, Scott Shenker, and Walter Willinger. Network topology generators: Degree-based vs. structural. In *Proceedings of SIGCOMM 2002, Pittsburgh, PA.*, 2002.

[22] Duncan J. Watts. *Small Worlds: The Dynamics of Networks Between Order and Randomness.* Princeton University Press, 1999.

[23] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[24] Bernard M. Waxman. Routing of mulitpoint connections. *IEEE Journal on Selected Areas in Communications*, 6(9):1617–1622, December 1988.

[25] Ellen W. Zegura, Kenneth L. Calvert, and Michael J. Donahoo. A quantitative comparison of graph-based models for internet topology. *IEEE/ACM Transactions on Networking*, 5(6):770–783, 1997.

Vita

Christine Leigh Myers Nickel was born on May 21, 1975, in Fort Belvoir, Virginia. Her father was an officer in the United States Army and she lived in 4 countries, on 3 continents, before graduating from Red Land High School in Lewisberry, Pennsylvania. She attended the University of Alabama in Huntsville, receiving a Bachelor of Science in Mathematical Sciences in 1997. During her undergraduate years, she worked as a researcher for professors in the physics and chemistry departments. After graduation, she married Lee Alan Nickel, a chemical engineer from Brookville, Ohio.

Christine then pursued graduate studies in the Mathematical Sciences at the University of Alabama in Huntsville, receiving her Master of Science in 1999. During her two years in graduate school at UAH, she worked as a teaching assistant and was primary instructor for courses in Basic Algebra and Precalculus.

Christine and her husband then moved north to the Washington, DC area. She spent one year teaching for Northern Virginia Community College before returning to graduate studies at the Johns Hopkins University. In 2002, she received a Master of Science in Engineering in Mathematical Sciences. During her time at Johns Hopkins she has worked as both a teaching and research assistant. She assisted with courses in Discrete Mathematics, Graph Theory, and Cryptology and Coding, and co-instructed a Special Topics in Secondary Education. Additionally, in 2001 she spent several months interning at Systems Planning and Analysis. Currently, she lives in Landover Hills, Maryland, with her husband and two year old daughter, Alexandra. She is expecting her next child in February, 2007.